CEFET/RJ - CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA
CELSO SUCKOW DA FONSECA

Aprendizado de máquina para previsão imediata de precipitação com dados de satélite meteorológico

Caio Abreu Ribeiro, Lucas Moura Alcantara, Luis Henrique Soares dos Santos

Prof. Orientador:

Eduardo Bezerra, D.Sc.

Rio de Janeiro,
Abril de 2025

CEFET/RJ - CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA CELSO SUCKOW DA FONSECA

Aprendizado de máquina para previsão imediata de precipitação com dados de satélite meteorológico

Caio Abreu Ribeiro, Lucas Moura Alcantara, Luis Henrique Soares dos Santos

Projeto final apresentado em cumprimento às normas do Departamento de Educação Superior do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação.

Prof. Orientador:
Eduardo Bezerra, D.Sc.

Rio de Janeiro,
Abril de 2025

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

R484 Ribeiro, Caio Abreu.
Aprendizado de máquina para previsão imediata de precipitação com dados de satélite meteorológico / Caio Abreu Ribeiro, Lucas Moura Alcantara [e] Luis Henrique Soares dos Santos. — 2024.
xi, 57f. : il. ; enc.

Projeto Final (Graduação) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2024.
Bibliografia: f. 55-57
Orientador: Eduardo Bezerra

1. Computação. 2. Inteligência Artificial. 3. Aprendizado de máquina. 4. Meteorologia. I. Alcantara, Lucas Moura. II. Santos, Luis Henrique Soares dos. III. Bezerra, Eduardo. (Orient.). IV. Título.

CDD 004

Elaborada pela bibliotecária Tania Mello – CRB/7 nº 5507/04

DEDICATÓRIA

A todos que nos apoiaram durante esta jornada.
Aos professores pelos ensinamentos e
orientações e ao CEFET/RJ, por nos
proporcionar esta experiência incrível.

AGRADECIMENTOS

Gostaríamos de agradecer ao CEFET/RJ pela oportunidade, ao professor Eduardo Bezerra como nosso orientador e professor. Deixamos também nossos mais sinceros agradecimentos a todos que nos apoiaram durante essa caminhada.

RESUMO

A utilização de técnicas de Aprendizado de Máquina tem sido cada vez mais comum no auxílio a diversos aspectos do cotidiano da população, incluindo a precisão da previsão meteorológica. Com a utilização dessas técnicas, o modelo preditivo pode aprender com dados históricos e gerar previsões a partir de novas informações, o que é fundamental para situações de risco como enchentes, deslizamentos de terra e chuvas fortes. A previsão imediata (nowcasting) é um tipo de método que envolve a análise em tempo real das condições meteorológicas, com o propósito de prever eventos futuros relevantes (com horizonte de tempo em torno de 1 até 6 horas). Esse método é fundamental no contexto de segurança e mobilidade urbana e para possibilitar tomadas de decisão mais rápidas e eficientes. Dessa forma, métodos de previsão imediata podem ajudar a população a se preparar melhor para situações climáticas adversas e reduzir os riscos de desastres naturais. O objetivo deste estudo é empregar informações do satélite GOES 16, um avançado satélite meteorológico desenvolvido pela NOAA (Administração Nacional Oceânica e Atmosférica dos EUA), combinado com técnicas modernas de previsão imediata, para criar previsões de curto prazo sobre chuvas e precipitações. Com a aplicação de técnicas de Aprendizado de Máquina, nosso propósito é gerar modelos preditivos capazes de considerar uma grande quantidade de variáveis e informações.

Palavras-chaves: Inteligência Artificial; Aprendizado de Máquina; Meteorologia; Previsão Imediata

ABSTRACT

The use of Machine Learning techniques has become increasingly common in assisting various aspects of daily life for the population, including the accuracy of weather forecasting. By utilizing these techniques, the predictive model can learn from historical data and generate forecasts based on new information, which is crucial for risk situations such as floods, landslides, and heavy rainfall. Nowcasting is a type of method that involves real-time analysis of weather conditions with the purpose of predicting relevant future events (with a time horizon ranging from 1 to 6 hours). This method is essential in the context of safety and urban mobility and enables faster and more efficient decision-making. Thus, nowcasting methods can help the population better prepare for adverse weather situations and reduce the risks of natural disasters. In this context, the present work aims to use data from the GOES-16 satellite, which is one of the latest meteorological satellites launched by the National Oceanic and Atmospheric Administration (NOAA) of the United States, in conjunction with these technologies to generate short-term precipitation weather forecasts. By applying Machine Learning techniques, our purpose is to generate predictive models capable of considering a large amount of variables and information.

Keywords: Artificial Intelligence; Machine Learning; Meteorological Prediction; Nowcasting

Sumário

1 Introdução **1**

1.1 Contextualização 1

1.2 Motivação 3

1.3 Objetivos 4

1.4 Metodologia 5

1.5 Organização dos capítulos 5

2 Fundamentação Teórica **6**

2.1 Meteorologia 6

2.1.1 Previsão de curto prazo 6

2.1.2 Meteorologia por meio de satélites 7

2.2 Satélite GOES-16 10

2.2.1 Produtos 11

2.3 API AWS para acesso aos dados do GOES-16 15

2.3.1 Acesso programático 15

2.3.2 Formato NetCDF 16

2.3.3 Exemplo 17

2.4 Redes neurais convolucionais unidimensionais 19

3 Trabalhos relacionados **22**

3.1 Seleção dos trabalhos 22

3.1.1 Critérios de inclusão 22

3.1.2 Critérios de exclusão 23

3.1.3 Critérios de priorização 23

3.2 Resultados 24

3.2.1 Comparação entre os trabalhos relacionados 24

4 Desenvolvimento **27**

4.1 Obtenção dos dados 27

4.1.1 INMET 27

4.1.2 GLM 28

4.1.3 TPW 29
4.2 Filtragem 31
4.3 Tratamento dos dados 31
4.4 Desbalanceamento dos dados 33
4.5 Integração 35
4.6 Treinamento 36
4.7 Divisão dos dados em treino, validação, teste 37
4.7.1 Classificação binária 37
4.7.2 Classificação ordinal 38
4.8 Avaliação dos modelos 39

5 Avaliação experimental 41
5.1 Conjunto de dados 41
5.2 Construção das redes neurais 42
5.2.1 Estrutura da Rede Neural 42
5.2.2 Hiperparâmetros 44
5.3 Resultados 45
5.3.1 Discussão 51

6 Considerações finais 53
6.1 Análise retrospectiva 53
6.2 Trabalhos futuros 54

Referências 54

Lista de Figuras

FIGURA 2.1: Satelite GOES 16, fonte: Wikipedia. 11

FIGURA 2.2: Esquema ilustrando o conceito de TPW, fonte: EUMeTrain. 12

FIGURA 2.3: Exemplo de imagem do produto TPW. 13

FIGURA 2.4: Exemplo de imagem do produto GLM — Foto: [NASA]. 15

FIGURA 2.5: Exemplo ilustrativo do acesso ao S3. 18

FIGURA 2.6: Exemplo de rede neural com três camadas de convolução unidimensionais. Retirada de Kiranyaz et al. [2021]. 20

FIGURA 4.1: Mapa do Rio de Janeiro e da região de interesse demarcada em vermelho. 31

FIGURA 5.1: Coeficiente de Matthews para cada combinação por estação. 46

FIGURA 5.2: F1-Score das diferentes combinações de dados para cada nível de chuva na Estação A602. 47

FIGURA 5.3: F1-Score das diferentes combinações de dados para cada nível de chuva na Estação A621. 48

FIGURA 5.4: F1-Score das diferentes combinações de dados para cada nível de chuva na Estação A627. 49

FIGURA 5.5: F1-Score das diferentes combinações de dados para cada nível de chuva na Estação A636. 50

FIGURA 5.6: F1-Score das diferentes combinações de dados para cada nível de chuva na Estação A652. 51

Lista de Tabelas

TABELA 4.1: Exemplo dataframe INMET. 27

TABELA 4.2: Explicação dos parâmetros do dataframe. 28

TABELA 4.3: Dataframe original. 30

TABELA 4.4: Dataframe após aplicação do metodo “mean()” 30

TABELA 4.5: Estações meteorológicas e seus nomes correspondentes. 34

TABELA 4.6: Percentual dos dados de cada classe em cada estação considerando a classificação ordinal. 34

TABELA 4.7: Percentual dos dados de cada classe em cada estação considerando a classificação binário. 34

TABELA 4.8: TPW dataframe antes de merge e criação da nova coluna. 35

TABELA 4.9: TPW dataframe depois do merge e criação da nova coluna. 36

TABELA 4.10: Níveis de precipitação definidos pelo alerta-rio. 38

TABELA 5.1: Estações meteorológicas e seus nomes correspondentes. 45

Lista de Abreviações

ABI	*Advanced Baseline Imager*	10, 12, 54
AM	*Aprendizado de Máquina*	2, 3, 4, 5, 19, 22, 25, 26
ANN	*Rede Neural Artificial*	6, 19, 41, 42
AWS	*Amazon Web Services*	15, 16, 18, 28, 29
CNN	*Redes Neurais Convolucionais*	2, 19
CNN 1D	*Redes Neurais Convolucionais Unidimensionais*	2, 5, 6, 19, 20, 21, 54
COR	*Centro de Operações Rio*	3
ECMWF	*European Centre for Medium-Range Weather Forecasts*	3
GLM	*Geostationary Lightning Mapper*	10, 14, 27, 28, 29, 32, 45, 46, 48, 49, 51, 52, 54
GOES	Geostationary Operational Environmental Satellites	8, 14
GOES-16	*Geostationary Operational Environmental Satellite-16*	3, 4, 5, 6, 10, 11, 14, 15, 16, 17, 26, 28, 29, 31, 45, 54
INMET	*Instituto Nacional de Meteorologia*	3, 27, 31, 35, 41, 45, 46, 48, 49, 50, 51
NOAA	*Administração Nacional Oceânica e Atmosférica*	5, 8, 10, 29
NWP	*Numerical Weather Prediction*	2
TPW	*Total Precipitable Water*	12, 13, 27, 29, 30, 32, 33, 35, 45, 46, 48, 49, 50, 52, 53, 54
UCAR	*University Corporation for Atmospheric Research*	16

Capítulo 1

Introdução

Neste capítulo de introdução apresentaremos a contextualização, motivação e objetivos do trabalho assim como sua metodologia e organização.

1.1 Contextualização

Devido às mudanças climáticas globais, eventos extremos como chuvas intensas tornaram-se mais frequentes. O aumento das concentrações de gases de efeito estufa na atmosfera está contribuindo significativamente para o aquecimento global, desencadeando alterações nos padrões climáticos. Como resultado, observamos a ocorrência mais frequente de chuvas intensas e imprevisíveis em várias partes do mundo.

Essas chuvas intensas têm o potencial de causar inundações, deslizamentos de terra e impactos socioeconômicos significativos. Regiões que antes eram consideradas relativamente seguras agora enfrentam uma maior suscetibilidade a riscos climáticos imprevistos, resultando em danos à infraestrutura, deslocamento de populações e interrupção das atividades do dia a dia.

Outro exemplo de mudança climática é o aumento das temperaturas globais. De acordo com Stern [2006] o aumento das temperaturas globais terá impactos significativos, como por exemplo: um aumento de 1°C poderia levar à diminuição de geleiras e à escassez de água para milhões de pessoas, exacerbando doenças climáticas. Atingindo 2°C, haveria queda na produção agrícola africana, crescimento da ameaça da malária e risco de extinção de espécies. Se as temperaturas subissem 3°C, o sul da Europa enfrentaria secas, até 4 bilhões de pessoas sofreriam com escassez de água e a fome se disseminaria, além de haver a possibilidade de colapso na Amazônia. Elevações de 4°C a 5°C resultariam em redução das safras africanas, ameaça às geleiras do Himalaia, impacto em ecossistemas marinhos e um aumento no perigo de inundações em cidades costeiras de grande importância [Marengo, 2008].

O termo *nowcasting* é usado para enfatizar que as previsões são para o período de tempo de 0-6 horas em uma escala espacial de não mais do que alguns quilômetros, com uma frequência de atualização de 1 hora ou menos. O *nowcasting* desempenha um papel eficaz em muitas ope-

rações táticas de suporte meteorológico, incluindo o clima da aviação, clima marítimo, clima de incêndios florestais, tempestades locais severas e enchentes [Wilson et al., 2010]. Embora tenha havido muito progresso nas últimas décadas o *nowcasting* continua sendo um desafio. Os métodos existentes de *nowcasting* podem ser classificados em três categorias: técnicas de extrapolação baseadas em dados de radar, modelos de *Numerical Weather Prediction* (NWP) e sistemas especialistas baseados em conhecimento que combinam NWP e técnicas de extrapolação [Han et al., 2017].

O *Aprendizado de Máquina* (AM) permite que computadores descubram conhecimentos que não foram programados explicitamente. O objetivo geralmente é prever uma variável alvo y, dadas muitas variáveis preditoras x_j. O AM tem muitas vantagens sobre o raciocínio humano e o NWP [Lagerquist et al., 2017]. O programa começa não só sem nenhum conhecimento de padrões específicos a serem inseridos, mas também sem nenhum operador para processar as entradas. Os operadores são gerados e refinados pelo próprio programa. Não só o programa aprende informações sobre diferentes padrões, mas também aprende ou constrói, pelo menos em parte, um código secundário apropriado para a análise do conjunto particular de padrões inserido nele [Uhr and Vossler, 1961].

Tecnologias de AM dão suporte a muitos aspectos da sociedade moderna: desde buscas na web até filtragem de conteúdo em redes sociais, recomendações em sites de comércio eletrônico, e cada vez mais presente em produtos de consumo, como câmeras e *smartphones*. Sistemas de aprendizado de máquina são usados para identificar objetos em imagens, transcrever fala em texto, associar notícias, *posts* ou produtos aos interesses dos usuários, e selecionar resultados relevantes de busca. Cada vez mais, essas aplicações fazem uso de uma classe de técnicas chamada *deep learning* [Lecun et al., 2015]. O *deep learning* é um dos métodos de aprendizado de máquina para encontrar características em dados enormes usando transformação não-linear. Agora é comumente usado para aprendizado supervisionado em muitas áreas. Em particular, as *Redes Neurais Convolucionais* (CNN) [Lee and Song, 2019].

O projeto AtmoSeer[1] consiste em um pipeline que disponibiliza a construção de modelos de previsão de chuva utilizando *Redes Neurais Convolucionais Unidimensionais* (CNN 1D). Esse pipeline é configurável com diversas fontes de dados meteorológicos. O pipeline do projeto é composto por três etapas: obtenção de dados, pré-processamento dos dados e treinamento do modelo. Cada uma dessas etapas é implementada por meio de programas escritos na linguagem

[1] `https://github.com/MLRG-CEFET-RJ/atmoseer/`

Python. No momento, o AtmoSeer utiliza como principais fontes de dados o *Instituto Nacional de Meteorologia* (INMET), o ERA5, produzida pela *European Centre for Medium-Range Weather Forecasts* (ECMWF) e o *Centro de Operações Rio* (COR), visando alcançar alta precisão nos modelos de previsão de curto prazo.

Outra fonte de dados que tem o potencial de aumentar o poder preditivos do modelos gerados pelo AtmoSeer são os satélites. Em particular, o *Geostationary Operational Environmental Satellite-16* (GOES-16) é um satélite meteorológico operado pela Administração Nacional Oceânica e Atmosférica (NOAA) dos Estados Unidos. Ele faz parte de uma nova geração de satélites geostacionários que fornecem observações e monitoramento contínuos da Terra.

O GOES-16 foi lançado em novembro de 2016 e está posicionado em uma órbita geoestacionária a cerca de 35.800 km acima do equador, cobrindo a região das Américas. Ele é equipado com uma variedade de instrumentos avançados que permitem capturar imagens e coletar dados sobre o clima, condições atmosféricas, tempestades, furacões, incêndios florestais e outros fenômenos meteorológicos em alta resolução e em tempo real.

Entre os instrumentos a bordo do GOES-16, destacam-se o Advanced Baseline Imager (ABI), que fornece imagens visíveis e infravermelhas em alta resolução espacial e temporal, e o Geostationary Lightning Mapper (GLM), que monitora a atividade de raios em toda a área de cobertura do satélite.

Com suas capacidades avançadas, o GOES-16 desempenha um papel crucial no monitoramento e previsão do tempo, permitindo aos meteorologistas e cientistas obter informações valiosas para tomar decisões informadas e alertar sobre condições climáticas perigosas.

1.2 Motivação

Embora a chuva tenha sido crucial para a sociedade desde seus primórdios, a previsão da precipitação ainda é um desafio complexo, pois depende de inúmeros fatores para que ocorra de maneira correta. Outro desafio que se apresenta é lidar com o grande número de datas disponíveis. Para otimizar o tempo neste trabalho, foi necessário buscar e utilizar um conjunto de datas previamente testadas.

Deste modo, o AM se apresenta como principal e mais interessante meio de se obter melhores resultados visto que possui a capacidade de lidar com a complexidade dos dados. O AM é capaz de processar e extrair padrões não triviais desses dados, permitindo uma modelagem robusta e, desta maneira, obter previsões mais acuradas sobre a ocorrência de chuva. Essa capa-

cidade de lidar com a complexidade dos dados é fundamental para obter *insights* significativos e auxiliar em tomadas de decisão informadas em diversas áreas, como agricultura, gestão de recursos hídricos e prevenção de desastres naturais.

Assim sendo, investigar outras fontes de dados para modelos de previsão se faz fundamental para melhorar a precisão, evitar vieses e erros, descobrir correlações e padrões ocultos, aumentar a adaptabilidade dos modelos e garantir a robustez do sistema de previsão. A diversificação das fontes de dados proporciona uma visão mais completa do problema em questão, resultando em *insights* mais precisos possibilitando decisões mais fundamentadas.

1.3 Objetivos

A pergunta de pesquisa central deste trabalho é: "A inclusão dos dados provenientes do satélite GOES-16 melhora a capacidade preditiva dos modelos de previsão?". Com base nessa questão, este projeto visa adicionar uma quarta fonte de dados aos modelos de predição, a fim de gerar previsões mais precisas de chuva por meio de modelos de aprendizado de máquina.

Como mencionado anteriormente, este projeto tem como objetivo testar a adição de uma nova fonte de dados aos modelos de predição do AtmoSeer e avaliar se há uma melhora na capacidade preditiva, com foco na cidade do Rio de Janeiro. O modelo escolhido para esse estudo inicial será o mesmo utilizado ao longo do projeto inicial, a Rede Neural Convolucional Unidimensional (CNN 1D), que pode ser usada em tempo real e não exige um poder computacional muito elevado.

Os objetivos específicos deste projeto estão citados abaixo:

- Tomar como ponto de partida os modelos de predição existentes e incorporar uma quarta fonte de dados para inferir previsões mais precisas de chuva por meio de modelos de aprendizado de máquina.
- Comparar os resultados obtidos ao usar o modelo de predição com os dados do satélite geoestacionário GOES-16.
- Testar modelos de AM para realizar a previsão da ocorrência de chuva.

1.4 Metodologia

Primeiramente será feita uma análise detalhada no código do AtmoSeer, a fim de se obter uma melhor compreensão para podermos realizar as adaptações necessárias para a incorporação da nova fonte de dados do satélite GOES-16.

Após esta primeira etapa de análise do código, estudaremos o satélite GOES-16, buscando compreender melhor seu funcionamento e seus dados.

Na terceira etapa será realizado uma análise exploratória nos dados do GOES-16 antes de iniciar a pesquisa de fato. Como se trata de uma nova fonte de dados e a mesma vem de um satélite geoestacionário lançado recentemente pela *Administração Nacional Oceânica e Atmosférica* (NOAA), investigaremos quais dados são trazidos por esta fonte e como eles podem servir ao modelo. Nesta tarefa estaremos também realizando uma análise cuidadosa para identificar, caso haja, comportamentos inesperados.

Após esta terceira etapa de análise dos dados obtidos, essas informações serão usadas para treinar diversos modelos preditivos para precipitação. Esses modelos utilizam de AM e CNN 1D. A fim de melhorar o poder preditivo, os resultados aqui obtidos serão comparados com os resultados das outras fontes de dados que foram usadas, nos modelos do projeto inicial.

Adiante, foi feito um estudo de comparação dos algoritmos de AM para a geração de modelos de *nowcasting*, com a finalidade de nos auxiliar na melhora das previsões.

1.5 Organização dos capítulos

O restante deste trabalho está organizado nos seguintes capítulos. O Capítulo 2 tem como objetivo fundamentar os conhecimentos apresentados nesta pesquisa. O Capítulo 3 contém os trabalhos relacionados. No Capítulo 4 está a parte de estudo dos modelos e desenvolvimento e produção dos algoritmos. O Capítulo 5 abrange a análise dos resultados obtidos na etapa de desenvolvimento. O Capítulo 6 apresenta as conclusões.

Capítulo 2

Fundamentação Teórica

Neste capítulo, exploraremos em detalhes as definições dos tópicos importantes para a pesquisa. Primeiramente nos concentramos em explicar alguns conceitos meteorológicos na Seção 2.1, na Seção 2.2 será abordado temas referentes ao funcionamento do satélite GOES-16. Depois passamos por uma explicação de como funciona o acesso aos dados do GOES-16, seu formato e um exemplo do código na Seção 2.3. Por fim, na Seção 2.4 passamos por uma explicação sobre *Rede Neural Artificial* (ANN) e CNN 1D.

2.1 Meteorologia

O INMET define a meteorologia como a ciência que estuda os fenômenos que ocorrem na atmosfera, e está relacionada ao estado físico, dinâmico e químico da atmosfera, as interações entre elas e a superfície terrestre subjacente.

A meteorologia básica, como o próprio nome sugere, proporciona uma compreensão simplificada dos fenômenos atmosféricos que ocorrem em nosso dia a dia. Por meio de observações dos principais elementos meteorológicos, como temperatura, velocidade e direção do vento, tipos e quantidade de nuvens, é possível obter uma noção adequada de como o tempo está se comportando em um determinado momento e local [INMET].

2.1.1 Previsão de curto prazo

Nowcasting é definido como a previsão do presente, do futuro imediato. Em meteorologia, o *nowcasting* se concentra na previsão do tempo para as próximas horas, utilizando dados meteorológicos em tempo real, como imagens de radar, satélite e leituras de sensores. Nesse processo, é fundamental utilizar informações atualizadas para realizar previsões de variáveis [Banbura et al., 2010].

A metodologia e as métricas de verificação do nowcast devem ser cuidadosamente selecionadas para fornecer informações relevantes aos usuários. Os responsáveis pelo desenvolvimento

da verificação do nowcast devem compreender as necessidades dos usuários, identificar métodos de verificação adequados, selecionar métricas e gráficos apropriados, coletar previsões e observações representativas, calcular métricas de verificação usando ferramentas disponíveis, apresentar os resultados de forma significativa para os usuários e realizar avaliações regulares da qualidade do nowcast para fornecer informações contínuas sobre seu desempenho [Schmid et al., 2019].

No *nowcasting*, a detecção precoce de situações propícias a alertas e a disseminação rápida dessas informações são aspectos essenciais. A utilização adequada dos produtos de *nowcasting* e uma resposta correta aos alertas podem ajudar a otimizar as medidas de proteção e reduzir perdas decorrentes de desastres. O crescimento populacional e econômico aumentou os riscos durante condições meteorológicas severas, assim como os impactos financeiros de tempestades intensas. Além disso, as informações meteorológicas se tornaram indispensáveis no dia a dia devido à modernização da sociedade. As exigências do *nowcasting* para o público em geral abrangem uma variedade ampla de expectativas, envolvendo fenômenos meteorológicos severos e diferentes elementos relevantes para a segurança, saúde, vida cotidiana, turismo e entretenimento [Schmid et al., 2019].

2.1.2 Meteorologia por meio de satélites

Diversas variáveis meteorológicas são monitoradas para compreender as condições atmosféricas. As mais comuns incluem a temperatura do ar em graus Celsius (ºC) e a umidade relativa do ar em porcentagem (%). Outras variáveis menos comuns são a pressão atmosférica medida em hectopascais (hPa) ou milibares (mb), a velocidade do vento em quilômetros por hora (km/h) ou metros por segundo (m/s), e a direção do vento expressa em graus [INMET].

Essas são apenas algumas das variáveis meteorológicas mais comuns, mas existem muitas outras que podem ser monitoradas para compreender melhor as condições atmosféricas.

Antes dos balões de ar quente em 1792, a única forma de obter uma visão aérea da Terra era subir em árvores ou escalar montanhas. Quando a câmera foi desenvolvida em meados do século XIX, as pessoas começaram a tirar fotografias aéreas a partir de balões de ar quente [CIMSS]. Desde o lançamento do Sputnik, pela URSS em outubro de 1957, os satélites têm se tornado indispensáveis na vida moderna. Eles possibilitam o envio e recebimento de imagens e notícias, a comunicação por meio da Internet e telefones de longa distância, além de impulsionar pesquisas científicas e a exploração espacial. [Silva Filho et al., 2008].

Ou seja, o que começou com um balão e câmera se tornaram, nos dias atuais, satélites que monitoraram os oceanos, o clima e praticamente tudo o que acontece entre a superfície da Terra e o topo da exosfera, além disso, registram desastres naturais, como incêndios, terremotos, etc [CIMSS].

Satélites meteorológicos capturam imagens das nuvens na Terra, permitindo a observação e previsão de eventos meteorológicos como frentes frias, geadas, furacões e ciclones. Essa previsão é vital para a segurança das pessoas. Além disso, os dados dos satélites meteorológicos contribuem para a compreensão e quantificação dos fenômenos ligados às mudanças climáticas [Florenzano, 2008].

Dados de satélite são necessários para fornecer uma imagem completa e contínua das condições atmosféricas. Desde 1975, os satélites Geostationary Operational Environmental Satellites (GOES) têm sido fundamentais na previsão da chegada de tempestades severas e são uma das bases da previsão do tempo [NOAA, b].

As melhores imagens para visualização dos fenômenos meteorológicos na América do Sul são geradas pelos satélites meteorológicos das séries GOES. Estes satélites, operados pela NOAA, encontram-se em uma posição estratégica que abrange praticamente todo o continente americano e grande parte dos oceanos Atlântico e Pacífico [Ferreira, 2006].

Meteorologia por Satélite é o estudo da atmosfera e dos oceanos da Terra usando dados de sensoriamento remoto obtidos por satélites em órbita. Os meteorologistas realizam observações da atmosfera utilizando dois métodos: diretos, que medem as propriedades do ar em contato direto com os instrumentos, e indiretos, que obtêm informações sem contato físico com a região da atmosfera em análise [CIMSS].

Meteorologia por Satélite utiliza diferentes tipos de satélites, como os em órbita polar e em órbita geoestacionária. Os satélites em órbita polar, também chamados de sol-síncronos, pois cruzam o equador sempre na mesma hora local, viajam de norte a sul, cobrindo toda a Terra e oferecendo imagens de alta resolução para observar fenômenos atmosféricos. Já os satélites em órbita geoestacionária permanecem fixos em relação ao planeta, fornecendo uma visão contínua de uma região específica e permitindo o monitoramento em tempo real de sistemas meteorológicos de grande escala [Florenzano, 2008].

Satélites fazem medições de forma indireta ao detectar radiação eletromagnética (ondas de energia) provenientes das superfícies abaixo. Essa radiação é a forma pela qual nosso planeta interage com o resto do universo, sendo a base para o sensoriamento remoto. Instrumentos

em satélites registram diferentes partes do espectro eletromagnético absorvidas, emitidas ou dispersadas pela atmosfera terrestre. Esses instrumentos fornecem informações valiosas sobre o nosso planeta e auxiliam em todos os aspectos da vida moderna, oferecendo ferramentas de investigação e análise para cientistas e educadores. Os dados dos satélites nos ajudam a monitorar e cuidar melhor do meio ambiente, incluindo lagos, oceanos e paisagens [CIMSS].

Existem vários satélites que fornecem dados observacionais gratuitos, além do GOES-16 (cujos dados são utilizados em nosso trabalho). Alguns desses satélites são operados por agências espaciais de diferentes países e organizações internacionais. Aqui estão alguns exemplos:

1. MODIS (Moderate Resolution Imaging Spectroradiometer): Os satélites Terra e Aqua da NASA carregam o instrumento MODIS, que fornece observações da Terra em várias bandas espectrais.

2. Sentinel Satellites: Os satélites Sentinel fazem parte do programa Copernicus da União Europeia. Eles incluem uma série de satélites com diferentes sensores para monitorar a Terra, incluindo Sentinel-1 (radar de abertura sintética), Sentinel-2 (imagens ópticas de alta resolução) e outros.

3. Landsat: A série de satélites Landsat, operada pela NASA e pelo Serviço Geológico dos Estados Unidos (USGS), fornece imagens de alta resolução temporal da superfície terrestre.

4. NOAA POES Series: A Administração Nacional Oceânica e Atmosférica (NOAA) opera satélites da série Polar-orbiting Operational Environmental Satellite (POES), que coletam dados meteorológicos e climáticos.

5. METOP: A Agência Espacial Europeia (ESA) opera a série de satélites METOP em parceria com a EUMETSAT para fornecer dados meteorológicos e climáticos.

6. Suomi NPP (National Polar-orbiting Partnership): Operado pela NASA e pela NOAA, o satélite Suomi NPP carrega instrumentos que fornecem dados sobre a atmosfera, oceanos e superfície terrestre.

7. Aqua Satellite: Além do Terra Satellite mencionado anteriormente, a NASA também opera o satélite Aqua, que carrega instrumentos semelhantes para observação da Terra.

2.2 Satélite GOES-16

O satélite GOES-16 (Geostationary Operational Environmental Satellite 16) representa uma conquista notável no campo da observação terrestre e meteorologia. Este satélite desempenha um papel vital na coleta de informações essenciais relacionadas ao clima e ao ambiente terrestre nas Américas. Uma de suas características fundamentais é sua órbita geossíncrona, que o mantém fixo em relação a um ponto na superfície da Terra, permitindo observar continuamente uma região específica do planeta [NOAA and NASA].

Os instrumentos a bordo do GOES-16 são projetados com tecnologia avançada para adquirir uma variedade de dados relacionados ao clima e ao ambiente. Um de seus principais instrumentos é o Advanced Baseline Imager (*Advanced Baseline Imager* (ABI)), que fornece imagens de alta resolução da Terra em diferentes comprimentos de onda. Essas imagens são cruciais para observar nuvens, tempestades, eventos climáticos extremos e outras condições ambientais em tempo real.

Outro instrumento importante é o Geostationary Lightning Mapper (*Geostationary Lightning Mapper* (GLM)), que detecta e mapeia relâmpagos na região de visualização do satélite. O monitoramento de relâmpagos é fundamental para a previsão de tempestades severas e para a segurança pública. Além disso, o GOES-16 possui o Advanced Space Environment Monitor (ASAP) e outros sensores que observam o ambiente espacial, incluindo atividades solares e radiação.

A coleta de dados pelo GOES-16 ocorre constantemente, e essas informações são transmitidas de volta à Terra por meio de um sistema de comunicação de alta frequência. As estações terrestres da NOAA e outras agências relacionadas recebem esses dados, que são então processados e transformados em produtos úteis. Esses produtos incluem imagens, mapas, informações espectrais, medições de temperatura, umidade, entre outros.

A disponibilização dos produtos gerados a partir dos dados do GOES-16 é essencial para aplicações práticas, como previsão do tempo, monitoramento climático, pesquisa científica e tomada de decisões. Eles são amplamente utilizados por agências meteorológicas, pesquisadores e o público em geral, sendo acessíveis por meio de diversos canais, como sites, sistemas de distribuição de dados, serviços de previsão do tempo e aplicativos móveis.

A Figura 2.1 mostra o satélite GOES-16, seus instrumentos e onde eles estão posicionados no satélite, como por exemplo o ABI e o GLM que estão posicionados um ao lado do outro.

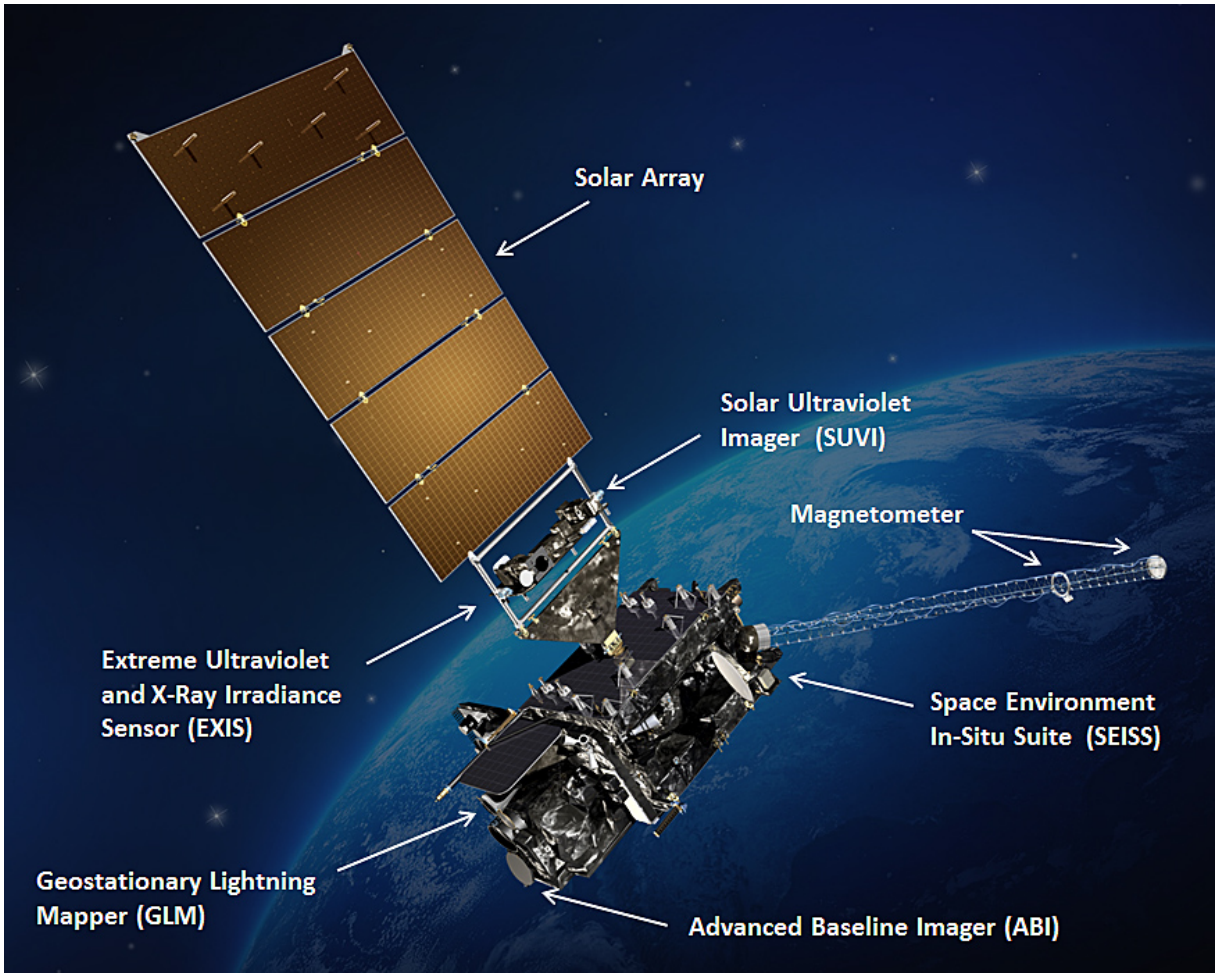

Figura 2.1: Satelite GOES 16, fonte: Wikipedia.

2.2.1 Produtos

O termo “produtos” se refere às informações e dados específicos que são derivados das observações e medições feitas pelo satélite. Esses produtos são gerados por meio de processamento e análise dos dados brutos coletados pelo satélite e são disponibilizados para cientistas, meteorologistas, agências governamentais e o público em geral.

Os produtos do GOES-16 podem incluir, por exemplo, imagens de satélite em diferentes faixas espectrais, como imagens visíveis, infravermelhas e de vapor de água. Essas imagens podem ser usadas para monitorar a formação de nuvens, identificar padrões climáticos, rastrear tempestades, monitorar a temperatura da superfície do mar e muitas outras aplicações.

Além das imagens, os produtos do GOES-16 também podem incluir dados de sensoriamento remoto, como informações sobre radiação solar, perfis verticais de temperatura e umidade, índices de estabilidade atmosférica, entre outros.

Total Precipitable Water

O *Total Precipitable Water* (TPW) busca medir a quantidade absoluta de água presente no ar, o que é diferente da umidade relativa, que está mais relacionada à capacidade do ar de reter água, influenciada pela temperatura [EUMeTrain].

A Figura 2.2 ilustra uma coluna vertical de ar que se estende desde o solo até o limite superior da atmosfera, com uma base de 1 metro quadrado. O TPW associado a esta coluna refere-se à quantidade de água que estaria presente caso todo o vapor de água contido nessa coluna se transformasse em forma líquida [EUMeTrain].

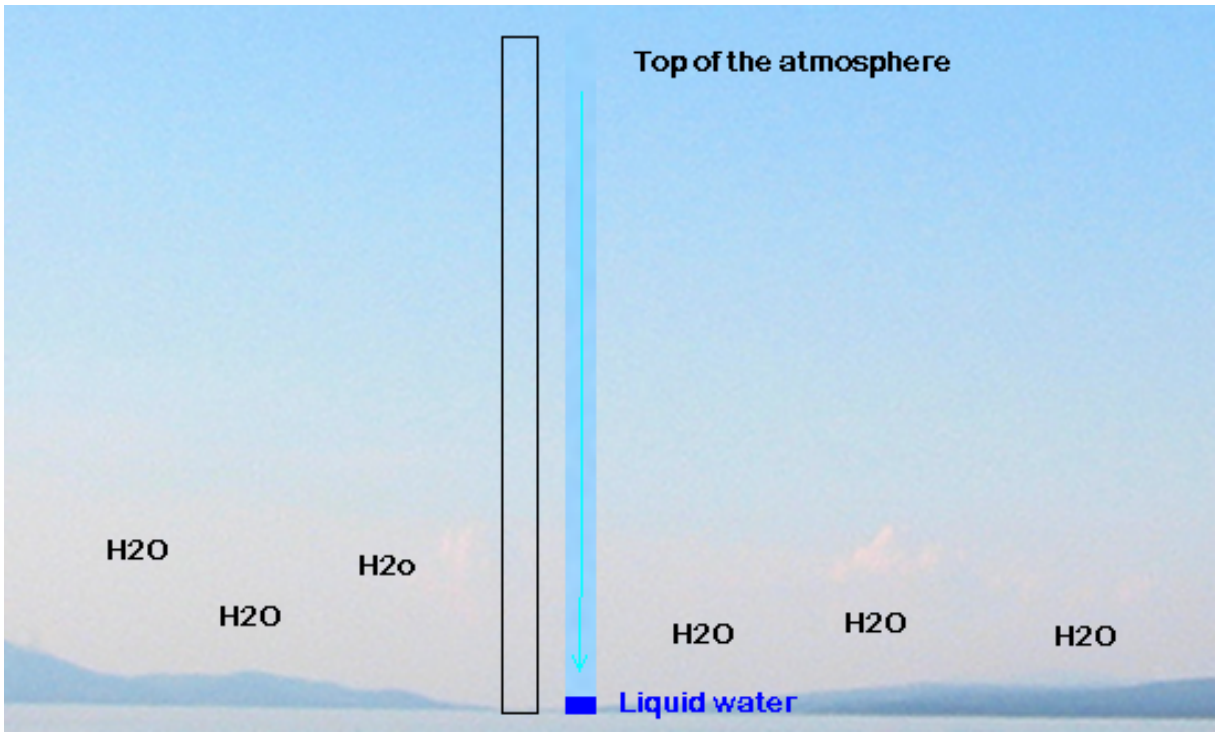

Figura 2.2: Esquema ilustrando o conceito de TPW, fonte: EUMeTrain.

Um dos produtos da ABI é o TPW. O produto é obtido através da análise dos perfis de umidade da atmosfera e reflete a quantidade total de umidade presente na coluna atmosférica desde o nível do solo até a parte mais alta da atmosfera. Essa informação é de grande utilidade para meteorologistas e hidrólogos, pois aprimora sua compreensão do cenário atual em diversas situações que exigem previsão de eventos, tais como chuvas intensas, enchentes repentinas, início do fluxo de ar do Golfo do México e início da monção no sudoeste dos Estados Unidos. Adicionalmente, o produto TPW é usado para inicializar o campo de umidade nos modelos numéricos de previsão do tempo, proporcionando informações cruciais para a precisão das previsões meteorológicas [NOAA, a].

A Figura 2.3 mostra um exemplo do TPW, que reflete a quantidade total de umidade presente na coluna atmosférica. A legenda abaixo mostra a quantidade de umidade em milímetros, de acordo com a cor. A captura dessa imagem ocorreu em 14 de dezembro de 2020.

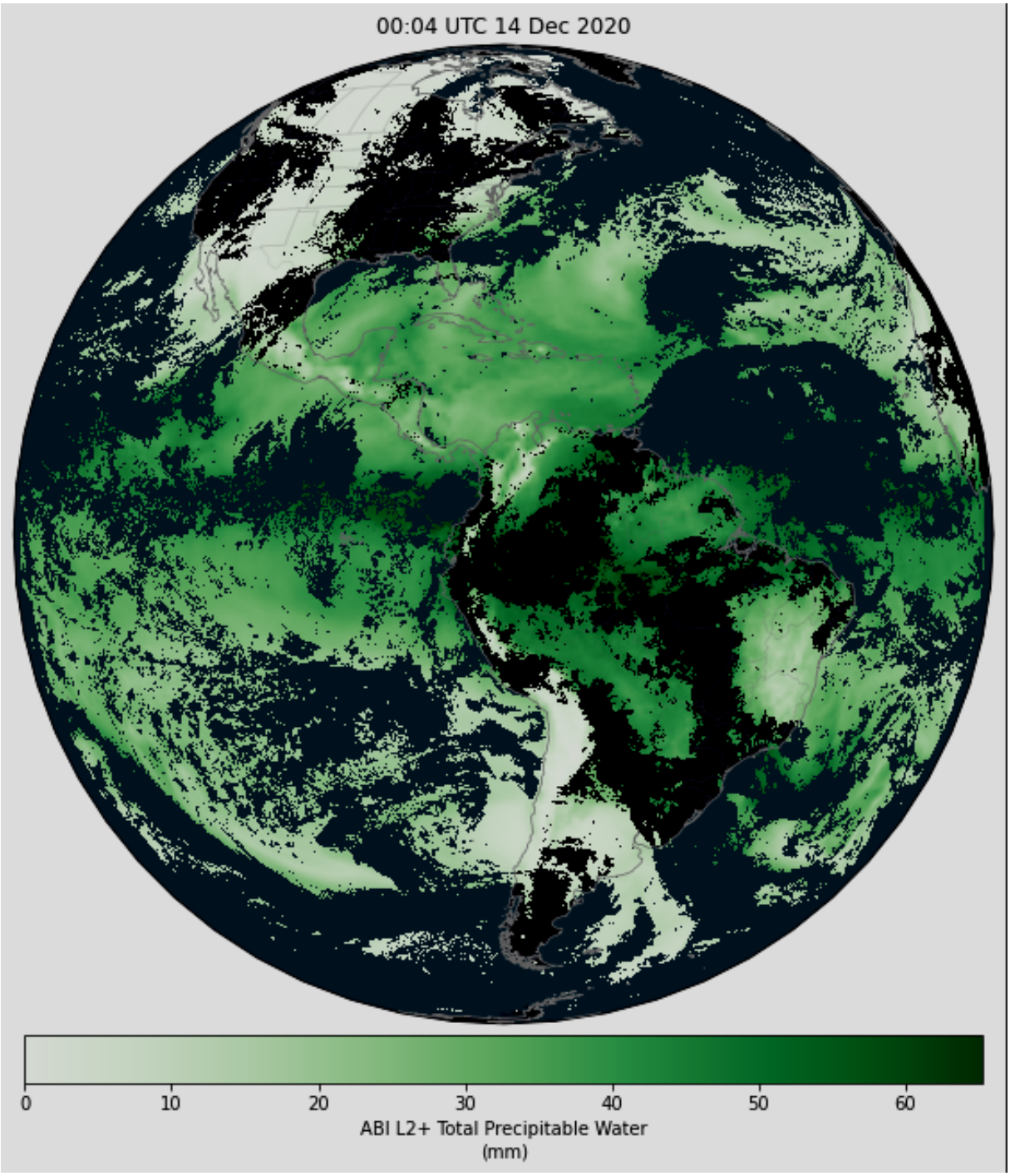

Figura 2.3: Exemplo de imagem do produto TPW.

Geostationary Lightning Mapper

O GLM é um dispositivo óptico instalado no satélite GOES-16 em órbita geoestacionária. Ele usa tecnologia de infravermelho próximo para detectar raios com alta precisão espacial e eficiência. O GLM começou a operar em março de 2017 e sua posição orbital permite monitorar continuamente uma região específica, como os Estados Unidos, detectando raios em uma taxa sem precedentes do espaço. Combinado com outros dados meteorológicos, como imagens de radar e nuvens, o GLM fornece informações em tempo real [NASA].

O GLM fornece informações cruciais para várias áreas, incluindo previsão do tempo, aviação, preparação para desastres e combate a incêndios. Suas medições melhoram os prazos de aviso antecipado de tornados e tempestades severas, alertando sobre os perigos de raios, aprimorando os modelos de previsão do tempo e facilitando o encaminhamento de aeronaves em regiões com poucas observações de tempestades. Além disso, o GLM melhora a monitorização de tempestades em áreas com cobertura de radar limitada, ajuda na detecção e previsão de chuvas intensas e enchentes repentinas, monitora a intensidade de ciclones tropicais e contribui para a climatologia de raios na região abrangida pelo satélite GOES [NOAA, a].

A Figura 2.4 apresenta um exemplo do GLM, a legenda ao lado mostra a quantidade de kilowatt-horas emitidos pelos relâmpagos, de acordo com a cor.

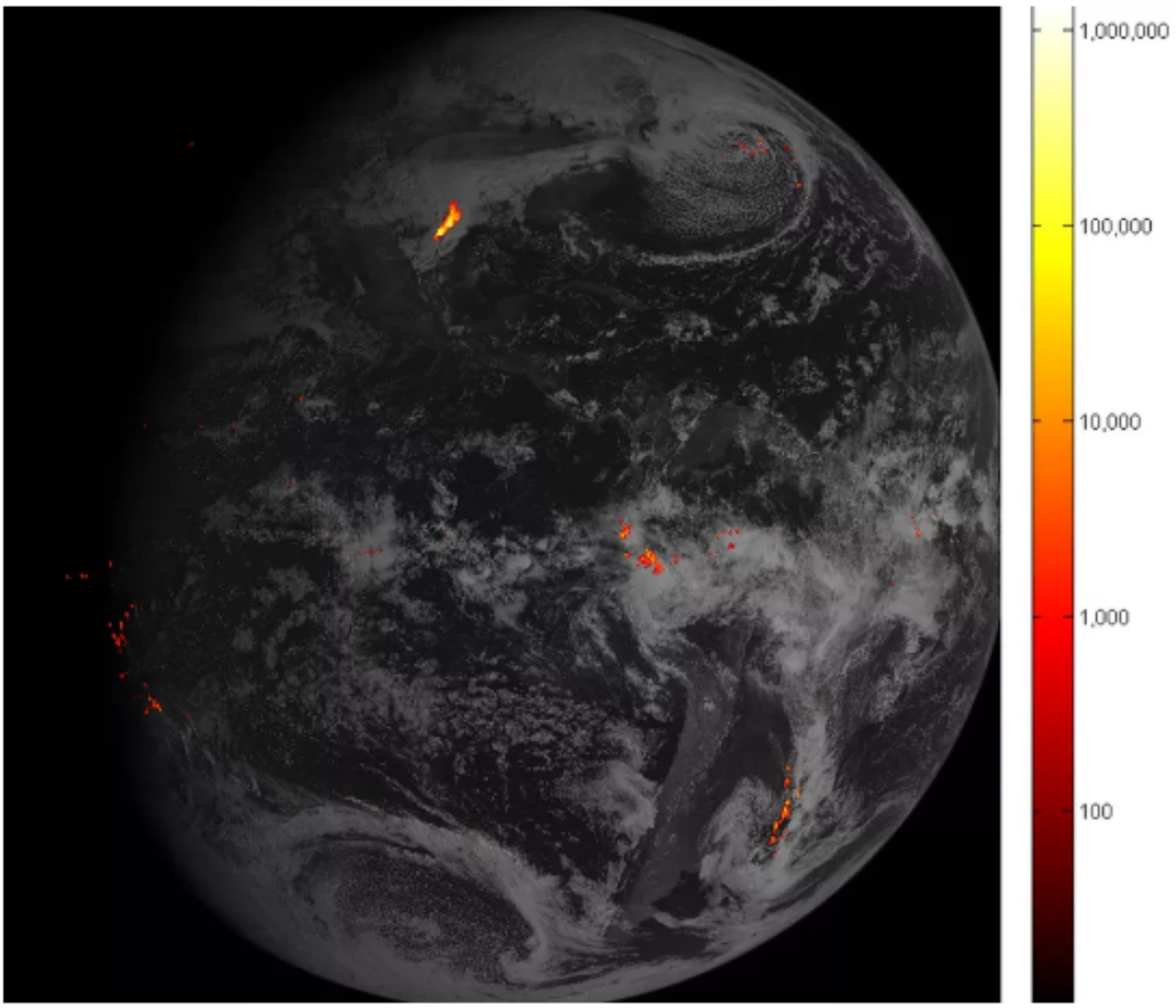

Figura 2.4: Exemplo de imagem do produto GLM — Foto: [NASA].

2.3 API AWS para acesso aos dados do GOES-16

A *Amazon Web Services* (AWS) fornece uma API chamada "Amazon S3 API" para acessar os dados do satélite GOES-16 armazenados no serviço de armazenamento em nuvem da Amazon, o Amazon S3 (Simple Storage Service). A API Amazon S3 permite realizar operações de leitura e escrita nos dados do GOES-16, bem como gerenciar o armazenamento e controle de acesso aos arquivos.

2.3.1 Acesso programático

Para acessar os dados do GOES-16 por meio da API AWS, é necessário seguir os seguintes passos:

1. **Criar uma conta na AWS**. É necessário criar uma para poder acessar os serviços e

recursos oferecidos.

2. **Configurar o acesso ao Amazon S3**. Após criar sua conta, é preciso configurar as credenciais de acesso para interagir com o Amazon S3. Isso envolve a criação de chaves de acesso (ID da chave de acesso e chave de acesso secreta) que serão usadas para autenticação.

3. **Instalar e configurar o SDK da AWS**. O próximo passo é instalar o SDK da AWS na máquina local. A AWS oferece SDKs para várias linguagens de programação, como Python, Java, Ruby, entre outras. É possível escolher o SDK compatível com a linguagem de programação a ser utilizada.

4. **Usar a API Amazon S3**. Com o SDK instalado e configurado, é possível utilizar as funções e métodos fornecidos pela API Amazon S3 para acessar os dados do GOES-16 armazenados no Amazon S3. Isso inclui listar os arquivos disponíveis, fazer o download dos arquivos desejados, gerenciar permissões de acesso, entre outras operações.

Os dados do satélite GOES-16 estão disponíveis em formatos específicos, como arquivos de imagem (por exemplo, formato NetCDF) e arquivos de dados meteorológicos. Dependendo do tipo de dados que devem ser acessados, pode ser necessário utilizar bibliotecas ou ferramentas adicionais para processar e interpretar esses arquivos de acordo com sua aplicação específica.

2.3.2 Formato NetCDF

NetCDF é um conjunto de bibliotecas de software e formatos de dados auto-descritivos e independentes de máquina que suportam a criação, acesso e compartilhamento de dados científicos orientados para arrays. O NetCDF foi desenvolvido e é mantido pela Unidata, que fornece ferramentas de software e dados para uso na educação e pesquisa em geociências. A Unidata faz parte dos Programas Comunitários da *University Corporation for Atmospheric Research* (UCAR). A Unidata é financiada principalmente pela National Science Foundation [Rew and Davis, 1990].

Os arquivos NetCDF contêm informações importantes sobre as condições climáticas em todo o continente americano, como imagens visuais e térmicas de alta resolução da Terra, medições da atividade solar e do ambiente espacial, dados sobre a atividade de relâmpagos e outras informações relevantes.

Esses dados são processados pelos instrumentos a bordo do satélite antes de serem armazenados em arquivos NetCDF. Ao utilizar o formato NetCDF, os cientistas e meteorologistas podem acessar e analisar esses dados de maneira eficiente, permitindo uma melhor compreensão do clima da Terra e uma previsão mais precisa do tempo.

Ao utilizar o formato NetCDF para armazenar esses dados, os cientistas podem acessar e analisar as informações coletadas pelo satélite de forma eficiente, permitindo uma melhor compreensão do clima da Terra e uma previsão mais precisa do tempo. A relação entre o formato NetCDF e o satélite GOES-16, portanto, é fundamental para a monitoração das condições climáticas da Terra e para a proteção das pessoas e do meio ambiente.

2.3.3 Exemplo

A Figura 2.5 apresenta um exemplo ilustrativo de uso da API Amazon S3 para acessar dados do satélite GOES-16. Este exemplo demonstra como usar a biblioteca **boto3** para criar um cliente S3, autenticar com as credenciais fornecidas e fazer o download de um arquivo específico do *bucket* especificado. Após o download, é possível processar os dados conforme necessário para suas análises ou aplicativos. Os mnemônicos 'ACCESS_KEY', 'SECRET_KEY', 'nome_do_bucket' e 'caminho/para/o/arquivo_no_s3' devem ser substituídos pelos valores corretos de acordo com as credenciais de acesso do usuário e o local onde os dados estão armazenados no Amazon S3.

```
import boto3

# Configurar as credenciais de acesso
access_key = 'ACCESS_KEY', secret_key = 'SECRET_KEY'

# Criar uma instância do cliente S3
s3_client = boto3.client('s3',
                aws_access_key_id=access_key,
                aws_secret_access_key=secret_key)

# Definir o nome do bucket e o
# caminho do objeto no Amazon S3
bucket_name = 'nome_do_bucket'
object_key = 'caminho/para/o/arquivo_no_s3'

# Fazer o download do arquivo do Amazon S3
response = s3_client.get_object(Bucket=bucket_name,
    Key=object_key)
data = response['Body'].read()

# Processar os dados do arquivo conforme necessário

# Exibir os dados
print(data)
```

Figura 2.5: Exemplo ilustrativo do acesso ao S3.

O código apresentado como exemplo utiliza a boto3, uma biblioteca Python desenvolvida pela AWS que fornece uma interface de programação de aplicativos (API) para interagir com serviços da AWS, como Amazon S3, Amazon EC2, Amazon DynamoDB e muitos outros. Ela permite que os desenvolvedores criem, configurem e gerenciem recursos da AWS usando código Python. Com a boto3, é possível realizar uma variedade de tarefas, como fazer o upload e o download de arquivos para o Amazon S3, criar e gerenciar instâncias do Amazon EC2, executar consultas no Amazon DynamoDB e acessar outros serviços e recursos da AWS. Essa biblioteca é amplamente utilizada por desenvolvedores e cientistas de dados que desejam automatizar tarefas na AWS, gerenciar infraestrutura em nuvem, interagir com serviços de armazenamento e banco de dados, implantar aplicativos e muito mais, tudo por meio de código Python.

Para utilizar a biblioteca boto3, é necessário ter credenciais válidas da AWS, como uma chave de acesso e uma chave secreta, para autenticar a API e acessar os serviços da AWS.

2.4 Redes neurais convolucionais unidimensionais

Uma ANN é uma estrutura composta por um número de neurônios interconectados. Cada unidade realiza um cálculo (geralmente não linear) na entrada, que pode ser também a saída de neurônios da camada anterior. O algoritmo mais importante que impulsiona o treinamento da ANN é o retropropagação (backpropagation) [Srinivasamurthy, 2018].

As ANNs enfrentam dois desafios principais. Um deles é a necessidade de usar redes neurais profundas para resolver problemas complexos, o que pode levar ao problema de "vanishing gradients". O outro desafio é selecionar as melhores características de uma entrada de alta dimensionalidade, o que pode levar a problemas de sobreajuste. As CNNs resolvem esse segundo problema de forma eficiente, lidando com a seleção de características. Além disso, as CNNs geralmente exigem menos neurônios, pois compartilham pesos com as saídas da convolução [Srinivasamurthy, 2018].

As CNNs estão inseridas no campo do Deep Learning, que é uma subcategoria do AM. As CNNs utilizam redes neurais artificiais com múltiplas camadas para alcançar alta precisão em tarefas como detecção de objetos, reconhecimento de fala, tradução de idiomas e outras [Srinivasamurthy, 2018].

Uma das principais motivações para a escolha da arquitetura CNN 1D é a sua capacidade de manter a consistência temporal dos dados. Por seus dados estarem em formato de vetor, esse tipo de modelo se prova mais vantajoso e até mesmo preferível para tarefas que dependam do tempo, como por exemplo na criação de modelos para previsão em séries temporais. Além disso, sua arquitetura mais simples permite uma menor complexidade computacional e facilita o treinamento e implementação desse tipo de modelo [Kiranyaz et al., 2021]. Em contrapartida o modelo mais popular para CNN's são as bidimensionais (2D), muito comuns em aplicações para processamento e identificação de padrões em imagens e vídeos, elas têm esse nome pois seus dados são apresentados no formato de uma matriz quadrada.

A Figura 2.6 apresenta um exemplo de rede neural convolucional unidimensional. Nessa imagem observamos uma entrada (sinal) unidimensional de tamanho 1000x1, indicando ser 1000 amostras em uma única dimensão. Nas Camadas Convolutivas (CNN Layers) observamos que a rede possui três camadas convolucionais consecutivas, cada uma com um propósito específico de extração de características. A primeira camada convolucional tem 24 filtros (também conhecidos como kernels) e uma largura de 960, o que sugere que cada filtro processa 960

pontos de dados de cada vez. Esta camada provavelmente expande a dimensão do sinal em 24 vezes, criando uma representação de alta dimensão. A segunda camada convolucional também possui 24 filtros, mas com uma largura reduzida de 200. Isso pode ser uma indicação de que a rede está começando a compactar a representação ou focar em características mais localizadas. A terceira camada convolucional tem uma largura ainda menor de 10, indicando um foco em características ainda mais refinadas ou um resumo das características extraídas anteriormente. As Camadas Totalmente Conectadas (MLP Layers) vêm logo após as camadas convolucionais. A primeira camada do MLP tem 24 neurônios. Cada neurônio nesta camada está conectado a todos os elementos da saída da última camada convolucional. Essa camada começa a integrar as características aprendidas pelas camadas convolucionais em uma representação mais abstrata. A segunda camada do MLP também tem 24 neurônios e provavelmente serve para refinar ainda mais a representação abstrata que será usada na saída. A Saída da rede é um vetor de 2x1, o que sugere que a rede está configurada para classificação ou regressão de duas classes/valores. Isso pode ser usado, por exemplo, para determinar a presença ou ausência de um certo padrão no sinal ou para fazer uma previsão binária.

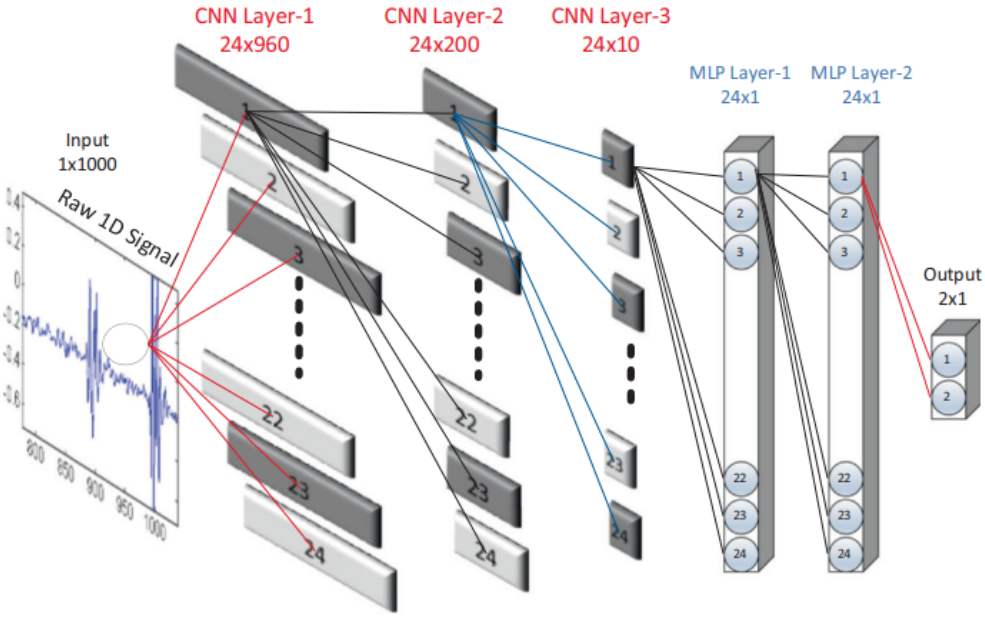

Figura 2.6: Exemplo de rede neural com três camadas de convolução unidimensionais. Retirada de Kiranyaz et al. [2021].

Uma camada de convolução de uma CNN 1D funciona recebendo um conjunto de dados no formato de um vetor e processando ele com o uso de uma estrutura (vetor) denominada "filtro" de tamanho pré definido. O filtro tem o objetivo de identificar padrões relevantes na

série temporal de entrada durante o treinamento.

O processamento é feito ao passar o filtro por cada um dos valores do vetor original e realizar uma operação matemática chamada convolução. A convolução pode ser representada pela seguinte equação, onde y é o vetor resultado da convolução, x é o vetor original e w é o vetor do filtro [e2eml, 2020]:

$$y_j = \sum_{k=-p}^{p} x_{j-k} w_k \tag{2.1}$$

A operação matemática citada acima é realizada para todo os valores do vetor original até que seja gerado um vetor novo e assim gerando o resultado da operação de convolução.

O processo de desenvolvimento de uma CNN 1D é iniciado na aquisição dos dados. Sendo a primeira etapa o pré-processamento, nela é necessário analisar e corrigir qualquer inconsistência nos dados, como valores faltantes ou fora do padrão estabelecido. Após isso determinamos uma janela de previsão, nesta etapa separamos os dados em três conjuntos distintos, os que serão usados para o treinamento do modelo, para a validação dos dados e por fim para testar a previsão. Com os dados separados podemos definir a rede neural e é nesse momento que a camada de convolução é estabelecida. Nessa camada os dados, que estão em formato de uma matriz unidimensional, serão processados através de um filtro e assim o resultado será passado para as próximas camadas de processamento da RNA. E por fim, antes da criação do modelo temos o treinamento, nele os dados já processados são passados pela rede neural de forma iterativa, com cada iteração o modelo se tornando mais preciso.

Capítulo 3

Trabalhos relacionados

Este capítulo tem como propósito discutir alguns projetos que apresentam semelhanças com esta pesquisa. Há um considerável número de artigos que utilizam algoritmos de AM e a dados de satélites. Veremos alguns deles aqui neste capítulo.

3.1 Seleção dos trabalhos

A busca por trabalhos relacionados foi conduzido com base em uma pesquisa abrangente que empregou as seguintes strings de busca: “machine learning”, “nowcasting” e “satellite data”. Essa combinação de termos-chave foi escolhida para identificar estudos relevantes que abordem a interseção entre os termos.

A string de busca foi aplicada na base de dados Scopus da Elsevier e Google Scholar, resultando em mais de 50 artigos, indicando uma considerável base de literatura que explora a aplicação de técnicas de aprendizado de máquina na análise de dados de satélite, especialmente no contexto do *nowcasting*.

No escopo, foram considerados artigos tanto na língua portuguesa quanto na inglesa e não houve nenhum tipo de restrição temporal.

- A base Scopus da Elsevier foi escolhida devido à extensa quantidade de material científico disponível em seu acervo, abrangendo também a área de pesquisa específica do presente trabalho.

3.1.1 Critérios de inclusão

Os critérios de inclusão têm como propósito orientar-se em relação ao tópico escolhido, ao mesmo tempo em que excluem pesquisas que não atendem aos objetivos do trabalho em questão. Esses critérios abrangem três categorias distintas de inclusão:

- **Critério 1 -** Estudos que abordem previsão imediata de precipitação utilizando dados de satélite meteorológico.

- **Critério 2 -** Trabalhos que, mesmo com dados de diferentes fontes ou variáveis meteorológicas, tenham o foco principal na previsão imediata de precipitação com aprendizado de máquina.
- **Critério 3 -** Estudos que demonstrem a aplicação de técnicas de aprendizado de máquina na análise de dados meteorológicos.

3.1.2 Critérios de exclusão

Os critérios de exclusão têm a finalidade de remover pesquisas que não atendem aos objetivos do presente trabalho.

- **Critério 1 -** Estudos que abordem fenômenos meteorológicos que não estão diretamente relacionados à precipitação, como, por exemplo, estudos sobre temperatura ou pressão atmosférica isoladamente.
- **Critério 2 -** Trabalhos que apresentem resultados inconsistentes ou que não forneçam evidências sólidas sobre a eficácia da abordagem de aprendizado de máquina na previsão imediata de precipitação com dados de satélite meteorológico.
- **Critério 3 -** Estudos que tenham uma abordagem predominantemente histórica, analisando somente eventos meteorológicos passados, em vez de se concentrar na previsão imediata com dados em tempo real.

3.1.3 Critérios de priorização

Após a inclusão dos artigos na busca, são aplicados critérios de prioridade para determinar a relevância do material resultante. Esses critérios foram categorizados em três níveis distintos, são eles:

- **Prioridade Alta -** Artigos que apresentam estratégias de aprendizado de máquina para a predição imediata de precipitação, utilizando dados de satélite meteorológico.
- **Prioridade Média -** Estudos que exploram métodos de previsão de precipitação, mesmo que não envolvam explicitamente o aprendizado de máquina, mas que façam uso de dados de satélite meteorológico.

- **Prioridade Baixa -** Artigos que não se concentram diretamente na predição imediata de precipitação, mas que mencionam variáveis potencialmente relevantes para a previsão e análise de dados de satélite meteorológico.

3.2 Resultados

Nenhum dos trabalhos que examinamos apresenta uma abordagem semelhante àquela proposta neste artigo, que seria a combinação de dados de estações meteorológicas terrestres com informações de satélite para desenvolver modelos de previsão imediata de precipitação. Entretanto foram encontrados artigos relevantes que serão citados a seguir.

3.2.1 Comparação entre os trabalhos relacionados

Nesse contexto, Bonelli and Marcacci [2008] apresenta um algoritmo desenvolvido para rastrear tempestades intensas usando dados de radares e relâmpagos. Ou seja, informações coletadas por um sistema de radar que emite ondas eletromagnéticas e as recebe de volta após refletirem em objetos na atmosfera, como nuvens e precipitação e dados que são coletados por uma rede de sensores que detectam a ocorrência de descargas elétricas na atmosfera durante uma tempestade. O algoritmo pode ser usado como um sistema de alerta para prevenir danos causados por eventos climáticos extremos e para estudar a correlação entre relâmpagos, precipitação e ocorrência de tornados. O artigo também descreve um método para avaliar a habilidade do algoritmo em prever a ocorrência de tempestades e apresenta um caso de estudo em que os dados foram usados para estudar a conexão entre relâmpagos de polaridade positiva e a formação de tornados. O algoritmo foi testado ao longo de uma temporada de tempestades e apresentou bom desempenho na previsão da ocorrência e movimento de tempestades. O estudo também destaca a importância de capturar e analisar dados de várias fontes para melhorar o *nowcasting* e reduzir os riscos associados a eventos climáticos extremos.

Kohn et al. [2011] busca aprimorar a previsão imediata (*nowcasting*) de tempestades severas por meio da análise de dados de raios. Utilizando a rede experimental de detecção de raios VLF ZEUS, composta por seis receptores de rádio distribuídos na Europa para identificar radiações eletromagnéticas de raios, e o software WDSS-II (Warning Decision Support System–Integrated Information) para auxiliar em decisões meteorológicas, foram conduzidas simulações de *nowcasting* ao longo de um ano na região do Mediterrâneo. O algoritmo de-

monstrou êxito na previsão da localização de aglomerados de raios, especialmente em eventos intensos, com taxas de detecção variando de 0,46 para nowcasts de 30 minutos a 0,25 para 120 minutos, e uma baixa taxa de alarme falso (0,03 para 30 minutos). O método foi implementado no projeto FLASH para monitorar áreas de risco de tempestades, ressaltando a utilidade do *nowcasting* baseado em raios na prevenção de inundações e seu potencial para fornecer previsões em tempo real. O estudo propõe a continuação da pesquisa para desenvolver relações confiáveis entre raios e precipitação na região do Mediterrâneo.

Em Shi et al. [2015] é proposto um modelo de rede neural chamado Convolutional LSTM (ConvLSTM) para a previsão de precipitação, mais especificamente, para prever a intensidade da chuva em uma região local em um curto período de tempo. A conclusão do artigo é que o modelo ConvLSTM proposto é capaz de prever com precisão a intensidade da chuva em uma região local em um curto período de tempo, superando outros métodos com os quais foi comparado.

Karagiannidis et al. [2016] traz uma ferramenta de previsão eficiente para atividade de raios na Grécia, usando imagens de infravermelho do MSG (Meteosat Second Generation),um satélite geostacionário, e dados de raios da rede ZEUS, que fornece dados em tempo real sobre atividade de raios. A ferramenta demonstra cerca de 80% de precisão em previsões durante 20 dias de teste, destacando sua utilidade para uso operacional.

Farnell et al. [2017] mostra um algoritmo de "Lightning Jump" que utiliza dados de raios para prever eventos climáticos severos, dispensando a necessidade de dados de radar. O "Lightning Jump" refere-se a aumentos súbitos na taxa total de raios antes de eventos como granizo e rajadas de vento. Testado na Catalunha de 2009 a 2014, o algoritmo mostrou eficácia na detecção desses eventos, ressaltando sua utilidade em países com orçamentos limitados para a previsão precoce da maioria dos eventos climáticos severos.

Agrawal et al. [2019] apresenta a aplicação de técnicas de *Deep Learning* para o problema de previsão de precipitação em curto prazo a partir de imagens. Fornece, especificamente, três classificações binárias que indicam se a taxa excede limites que correspondem aproximadamente a chuva traço, chuva leve e chuva moderada. É concluído com o levantamento de questões acerca do desempenho do AM comparado com métodos numéricos tradicionais.

Zhou et al. [2020] apresenta uma nova abordagem para a previsão de raios baseada em *Deep Learning* que integra dados de observação de várias fontes. A rede neural desenvolvida, chamada "LightningNet", é capaz de prever a ocorrência de raios com alta precisão e antece-

dência de até 30 minutos. Os dados utilizados incluem imagens de satélite, dados de radar e informações de localização de raios. Os resultados da validação mostram que a abordagem proposta supera outras técnicas de previsão de raios existentes. A aplicação prática dessa abordagem pode ajudar a prevenir danos causados por raios e melhorar a segurança em atividades ao ar livre.

Nespoli et al. [2022] tem como objetivo a previsão precisa da irradiância solar quinze minutos à frente, utilizando técnicas de Machine Learning, imagens de satélite e dados meteorológicos. Introduz-se um novo modelo projetado para detectar em tempo real nuvens que possam interferir nos raios solares direcionados a um ponto alvo, com o intuito de reduzir o esforço de treinamento de técnicas de aprendizado de máquina sem comprometer a precisão da previsão.

Leinonen et al. [2022] traz como objetivo fornecer uma avaliação sistemática do valor de várias fontes de dados para a previsão imediata de perigos causados por tempestades elétricas, utilizando uma abordagem de AM e dados provenientes de satélites, incluindo o GOES-16.

Uma conclusão comum que pode ser retirada destas pesquisas é o reconhecimento de que a tecnologia e o uso de AM para previsões de curto prazo ainda podem ser aprimorados e continuarão a se desenvolver ao longo do tempo, assim como a utilização de mais de uma fonte de dados podem vir a ser importantes para validar e aprimorar resultados.

Capítulo 4

Desenvolvimento

Este capítulo aborda a metodologia utilizada neste trabalho. A Seção 4.1 apresenta uma explicação sobre a obtenção dos dados das diferentes fontes utilizadas (INMET, GLM, TPW. Adiante, temos a Seção 4.2 que trata da descrição do filtro de região geográfica utilizado. A Seção 4.3 explica o processo de tratamento dos dados utilizados. Na Seção 4.5 temos a explicação de como é feita a integração dos dados. A explicação de como foi feito o treinamento do modelo está presente na Seção 4.6. A Seção 4.7 mostra como foi feita a divisão do dados entre treino, validação e teste. Por fim, a Seção 4.8 traz a avaliação dos modelos utilizados no trabalho.

4.1 Obtenção dos dados

4.1.1 INMET

O processo de obtenção dos dados do INMET inclui a identificação da estação meteorológica desejada e a definição do intervalo de anos. Para adquirir os dados é usado a própria API do INMET com base nessas informações, recupera os dados meteorológicos em formato JSON e é processado usando a biblioteca pandas. Por fim, salva as observações em um arquivo CSV, nomeado com base no código da estação. A resolução temporal do INMET é uma hora.

A Tabela 4.1 apresenta um exemplo dos dados adquiridos da API do INMET.

datetime	temperature	barometric_pressure	relative_humidity	wind_direction_u	wind_direction_v	hour_sin	hour_cos	precipitation
2018-01-01 00:00:00	0.534483	0.282132	0.724138	0.440824	0.289500	0.500000	1.000000	0.0
2018-01-01 01:00:00	0.525862	0.288401	0.735632	0.495047	0.356405	0.629410	0.982963	0.0
...	...	...	...	...	...	...	...	...
2023-11-24 13:00:00	0.383621	0.545455	0.632184	0.459594	0.505527	0.370590	0.017037	0.0

Tabela 4.1: Exemplo dataframe INMET.

Parâmetro	Descrição
temperature	Esta variável representa a temperatura registrada na estação meteorológica em um determinado momento. Parece estar normalizada para um valor entre 0 e 1, onde 0 corresponde a uma temperatura mínima e 1 a uma temperatura máxima.
barometric pressure	Essa variável indica a pressão barométrica medida na estação meteorológica. Novamente, parece estar normalizada entre 0 e 1, onde 0 pode representar uma pressão barométrica baixa e 1 uma pressão barométrica alta.
relative_humidity	Representa a umidade relativa do ar medida na estação meteorológica. A normalização entre 0 e 1 sugere que 0 pode indicar uma umidade baixa, enquanto 1 pode indicar uma umidade alta.
wind_direction_u wind_direction_v	Essas duas variáveis representam as componentes da direção do vento. A direção do vento é frequentemente representada em coordenadas polares, onde 'U' representa a componente horizontal do vento na direção leste-oeste (ou longitude) e 'V' representa a componente horizontal do vento na direção norte-sul (ou latitude).
precipitation	Esta variável indica a quantidade de precipitação (como chuva) registrada na estação meteorológica em um determinado momento. Parece não estar normalizada e pode ser medida em milímetros ou outra unidade de medida padrão de precipitação.
hour_sin hour_cos	Essas variáveis representam a hora do dia em forma de valores de seno e cosseno. Isso é comum em análises de séries temporais para levar em consideração a variação sazonal ao longo do dia. Elas podem ser usadas para capturar padrões sazonais nas medições.

Tabela 4.2: Explicação dos parâmetros do dataframe.

4.1.2 GLM

Utilizando os dados disponibilizados gratuitamente pela AWS, por meio de um sistema de armazenamento chamado buckets, assim é obtido todos os dados do satélite referentes a produtos específicos em um determinado intervalo de tempo. Os buckets facilitam o acesso e a coleta dos dados desejados, permitindo uma análise mais precisa e detalhada ao longo do tempo.

O formato do dado do produto GLM do satélite GOES-16 é baseado em dados binários. Os dados são organizados em uma estrutura específica que permite armazenar e transmitir informações relacionadas a eventos de relâmpagos capturados pelo instrumento GLM a bordo do satélite.

A resolução temporal do GLM é extremamente alta, pois ele é capaz de capturar eventos de relâmpagos em tempo real. Os dados são disponibilizados com uma taxa de atualização de cerca de 8 a 10 segundos, o que significa que novas informações são recebidas e processadas em intervalos muito curtos.

Em termos de resolução espacial, o GLM é capaz de fornecer informações precisas sobre a localização dos eventos de relâmpagos. Ele pode identificar a latitude e a longitude exatas dos pontos onde os relâmpagos ocorrem, com uma precisão de aproximadamente 10 km.

As principais variáveis que são medidas pelo GLM são a energia radiante dos relâmpagos e a taxa de repetição dos eventos. A energia radiante é uma medida da quantidade de energia emitida pelos relâmpagos em um determinado intervalo de tempo. A taxa de repetição se refere ao número de relâmpagos que ocorrem em uma área específica durante um determinado período.

Quanto à disponibilidade de informações sobre o produto GLM do satélite GOES-16, os dados começaram a ser coletados e disponibilizados a partir de 2017. O satélite GOES-16, operado pela NOAA dos Estados Unidos, é o primeiro satélite da série GOES-R (Geostationary Operational Environmental Satellite-R). O GLM é um dos instrumentos a bordo do satélite GOES-16 e tem como objetivo principal monitorar e fornecer informações em tempo real sobre atividades de relâmpagos nas Américas. Desde o seu lançamento e início das operações, o GLM tem fornecido dados valiosos para monitoramento de tempestades e previsão meteorológica em tempo real.

4.1.3 TPW

Assim como o GLM, os dados do TPW foram obtidos por meio do acesso gratuito aos dados fornecidos pela AWS, utilizando o sistema de armazenamento denominado “buckets”, desta forma, foi desenvolvido um script que possibilita a obtenção dos dados do satélite relacionados ao TPW em um intervalo de tempo definido.

O produto TPW é avaliado em todo o disco de observação do satélite GOES-16. Cada disco de observação é gerado a cada dez minutos. Coletamos cada varredura completa do produto TPW do período de 4 de dezembro de 2019 a 30 de setembro de 2023. Posteriormente, processamos cada varredura para extrair o valor associado à latitude e longitude da estação meteorológica de interesse. Como resultado, para a estação meteorológica em questão, obtemos uma série temporal univariada composta por valores medidos na localização geográfica correspondente à estação de interesse. Essa série temporal possui uma resolução temporal de 10

minutos, resultando em até 144 pontos diários.

Considerando que as séries temporais originárias das fontes de dados que estamos considerando apresentam resoluções temporais distintas, é necessário realizar uma conciliação para gerar uma série temporal combinada. Para realizar essa conciliação, agregamos a série temporal TPW, que possui uma resolução temporal mais detalhada. Especificamente, agrupamos cada seis observações consecutivas e calculamos a média correspondente. Por exemplo, para as observações realizadas às 10h10min, 10h20min, 10h30min, 10h40min, 10h50min e 11h00min, calculamos seus valores médios para gerar um único ponto de observação agregado às 11h00min. Dessa maneira, garantimos que o modelo não tenha acesso a informações futuras durante o treinamento.

A Tabela 4.4 mostra um exemplo do calculo da temperatura do TPW para os horários 12h00min e 13h00min. Onde o valor de 12h00min foi calculado com o valor médio de 12h00min, 12h15min e 12h45min enquanto o valor de 13h00min foi calculado somente com o valor de 13h30min.

datetime	**temperature**
2023-10-30 12:00:00	25.5
2023-10-30 12:15:00	26.2
2023-10-30 12:45:00	27.8
2023-10-30 13:30:00	28.5

Tabela 4.3: Dataframe original.

datetime	**temperature**
2023-10-30 12:00:00	26.5
2023-10-30 13:00:00	28.5

Tabela 4.4: Dataframe após aplicação do metodo “mean()”

Após a agregação descrita acima, temos duas séries temporais para uma determinada estação meteorológica de interesse, uma é unidimensional e a outra é tridimensional. Como ambas apresentam resolução temporal horária, podem ser combinadas de maneira trivial, resultando em uma série temporal de quatro dimensões (tratamento) para o ponto geográfico da estação meteorológica de interesse. Essa série temporal combinada pode, então, ser usada para ajustar um modelo de previsão imediata de precipitação para a proximidade da estação meteorológica de interesse.

4.2 Filtragem

Para obter os dados do GOES-16 filtrados por região, é necessário realizar um processo de filtragem inicial, limitando-os apenas à região do município do Rio de Janeiro. Isso permite uma análise mais precisa e direcionada às condições meteorológicas locais, contribuindo para a previsão de tempestades, o monitoramento e a tomada de decisões na região. Em seguida, é preciso fazer uma filtragem mais específica, selecionando apenas os dados da estação desejada.

A Figura 4.1 mostra o mapa da cidade do Rio de Janeiro e em vermelho está marcada a região de interesse utilizada para a filtragem dos dados.

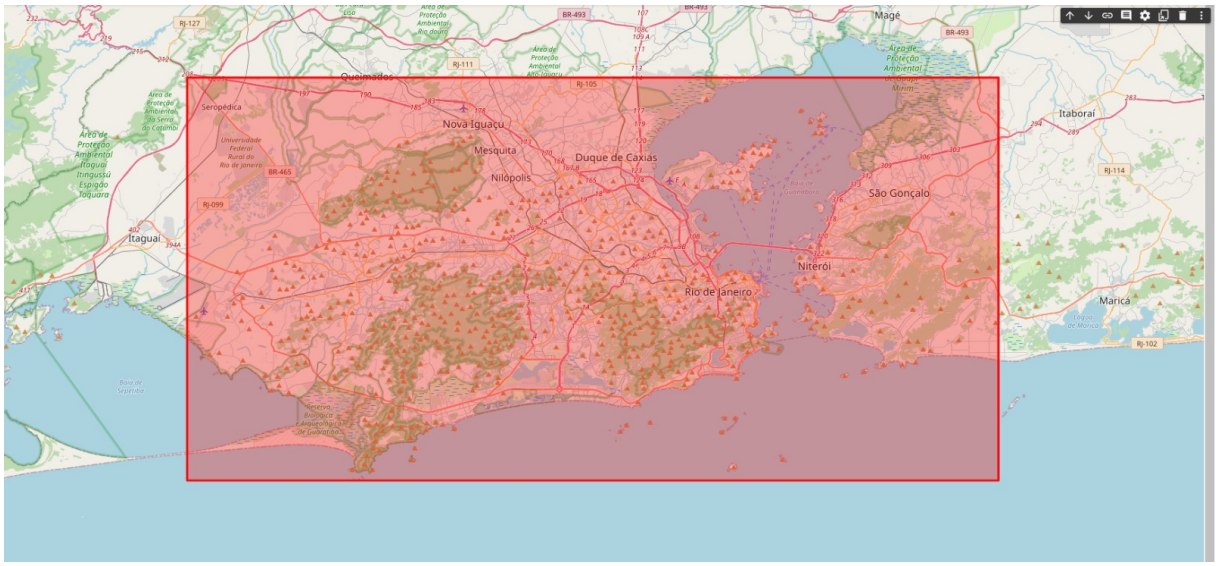

Figura 4.1: Mapa do Rio de Janeiro e da região de interesse demarcada em vermelho.

Para realizar esse filtro, utilizamos as coordenadas de latitude e longitude contidas nos dados baixados. Com base nessas informações, verificamos quais arquivos possuem eventos no Rio de Janeiro dentro do intervalo de tempo especificado. Dessa forma, conseguimos extrair apenas os dados relevantes para a análise desejada.

4.3 Tratamento dos dados

Para os dados do INMET, o processo de tratamento dos dados meteorológicos inicia-se com a carga dos dados brutos provenientes da estação meteorológica, utilizando a biblioteca Pandas. Uma vez carregados, são adicionados índices de data e hora aos dados, facilitando análises temporais. Em seguida, ocorre a normalização min-max dos dados, uma prática comum para garantir consistência nas escalas das variáveis. Para lidar com possíveis valores ausentes em

nosso conjunto de dados, optamos pelo uso do algoritmo KNNImputer, uma escolha estratégica para preservar a estrutura temporal e a integridade dos dados. O KNNImputer funciona identificando inicialmente as entradas ausentes, geralmente representadas por NaN ou outros marcadores. Uma vez identificados, o algoritmo calcula a distância (tipicamente a distância euclidiana, embora outras métricas possam ser aplicadas) entre o ponto de dado faltante e os outros pontos com dados completos. O processo segue selecionando os K vizinhos mais próximos - onde K é um número pré-definido - que são os pontos de dados mais semelhantes ao ponto com valores ausentes. A imputação é realizada com base nesses vizinhos, comumente através do cálculo da média ou mediana de seus valores. Essa metodologia é particularmente eficaz em conjuntos de dados multidimensionais, permitindo uma abordagem contextual e sofisticada para a imputação, ao contrário de métodos mais simplistas como a imputação pela média ou mediana geral. Ao utilizar o KNNImputer, garantimos que a imputação dos valores ausentes foi realizada de maneira a manter a consistência e a fidelidade dos padrões temporais presentes nos dados, essencial para a integridade de nossa análise. Além disso, são realizadas transformações adicionais, como a criação de características relacionadas ao vento e à hora, personalizando o conjunto de dados para análises específicas. O resultado final é armazenado em um arquivo parquet, pronto para análises subsequentes.

No tratamento dos dados relacionados ao produto GLM, o processo tem início com a navegação até o diretório contendo os arquivos NetCDF. Os arquivos são então coletados e processados em lotes de 1000 para otimização de memória. Durante o processamento, eventos de raios que não estão dentro das coordenadas geográficas específicas associadas à estação meteorológica são removidos. Os dados resultantes são convertidos em um formato tabular (DataFrame) e, em seguida, são armazenados em um arquivo Parquet. Esse método não apenas preserva a informação temporal dos eventos de raios, mas também otimiza o uso de recursos computacionais. O arquivo Parquet consolidado pode ser utilizado como uma fonte de dados pronta para análises detalhadas.

Em relação à série temporal TPW, observamos que essa fonte de dados apresenta um número significativo de valores ausentes. Em vez de simplesmente remover essas observações, aplicamos a técnica de usar uma coluna indicadora de valores ausentes. Uma coluna indicadora de valores ausentes é uma variável binária que indica se o valor correspondente em outra coluna está ausente ou não. É uma maneira de codificar a ausência de dados como uma característica. Para criar uma coluna indicadora de valores ausentes para os dados TPW, fizemos o

seguinte: para cada linha no conjunto de dados, verificamos se o valor na coluna TPW estava ausente. Se sim, definimos o valor na nova coluna como 1. Caso contrário, definimos o valor como 0. Uma vez que a coluna indicadora de valores ausentes é criada, utilizamos o método *fillna(method='bfill')*,o argumento *method='bfill'* é utilizado no pandas, uma biblioteca popular para análise de dados em Python. O objetivo do método fillna é preencher valores ausentes (NaN) em um DataFrame ou Series com valores específicos.

No caso do *method='bfill'*, "bfill" significa "backward fill" ou preenchimento para trás. Isso implica que os valores ausentes serão preenchidos com os valores que aparecem imediatamente depois deles na sequência de dados.

Realizamos diversas etapas de pré-processamento nas observações provenientes da estação meteorológica de interesse, conforme detalhado a seguir. Inicialmente, removemos observações em que a variável-alvo não estava definida. Posteriormente, aplicamos transformações padrão às variáveis cíclicas. Para a marca temporal (um valor originalmente no formato YYYY-MM-DD:HH:mm), procedemos à extração de seus componentes seno e cosseno. Assim, com base nos valores da hora e dos minutos extraídos do carimbo de data/hora, calculamos esses componentes utilizando a equação 4.1, equação 4.2 e equação 4.3:

$$horafloat = hora + minutos/60 \tag{4.1}$$

$$\text{senohora} = seno(2 \times \pi \times horafloat/24) \tag{4.2}$$

$$cossenohora = cosseno(2 \times \pi \times horafloat/24) \tag{4.3}$$

4.4 Desbalanceamento dos dados

Nesta seção, realizamos uma análise do desbalanceamento na distribuição de valores para cada classe nas cinco estações utilizadas nesse trabalho. A Tabela 4.5 apresenta o nome e código correspondente de cada estação meteorológica.

Primeiramente veremos, na tabela abaixo, o percentual de cada classe para a classificação ordinal.

Código da Estação	Nome da Estação
A602	Marambaia
A621	Vila militar
A627	Niterói
A636	Jacarepaguá
A652	Forte de Copacabana

Tabela 4.5: Estações meteorológicas e seus nomes correspondentes.

Class	A602	A621	A627	A636	A652
% Sem chuva (0)	$\approx 91,5$	$\approx 92,5$	$\approx 92,3$	$\approx 90,7$	$\approx 92,6$
% Chuva fraca (1)	$\approx 8,0$	$\approx 6,9$	$\approx 7,2$	$\approx 8,5$	$\approx 6,7$
% Chuva moderada (2)	<0,5	<0,5	<0,5	<0,7	<0,6
% Chuva forte (3)	<0,00	<0,00	<0,00	<0,00	<0,00
% Chuva extrema (4)	<0,00	<0,00	<0,00	<0,00	<0,00

Tabela 4.6: Percentual dos dados de cada classe em cada estação considerando a classificação ordinal.

Como visto na tabela 4.9, as classes “Chuva moderada”, “Chuva forte” e “Chuva extrema” possuem um baixo percentual na quantidade de dados, destacando-se as classes “Chuva forte” e “Chuva extrema”, ambas com percentuais inferiores a $0,00\%$ em todas as cinco estações. Isso faz com que um modelo treinado em cima dessa base de dados tenha dificuldade em aprender esses padrões menos frequentes e fique mais “viciado” nos que têm maior ocorrência.

Na tabela 4.7 está o percentual da quantidade de valores para as classes “Sem chuva” e “Chuva” para a classificação binário.

Class	A602	A621	A627	A636	A652
% Sem chuva (0)	91,5	92,5	92,3	90,7	92,6
% Chuva (1)	8,5	7,5	7,7	9,3	7,4

Tabela 4.7: Percentual dos dados de cada classe em cada estação considerando a classificação binário.

Assim como para a classificação ordinal, notamos que os dados também se encontram muito desbalanceados quando olhamos para o caso binário, isto é, de chuva ou não chuva. Isso também faz com que o modelo tenha mais dificuldade em aprender esses padrões na hora do treinamento da rede neural.

4.5 Integração

É importante manter um padrão temporal ao utilizar diferentes fontes de dados, pois cada fonte pode ter sua própria resolução temporal. Isso pode causar problemas ao tentar combinar diferentes séries temporais. Um exemplo disso são as estações meteorológicas do INMET, que coletam dados a cada hora.

Para padronizar o período de tempo em relação aos dados fornecidos pelo INMET, foi decidido usar uma resolução temporal de uma hora para os dados a serem usados para treinamento dos modelos. A partir do cálculo da média dos eventos de energia que aconteceu em cada hora, onde quando ocorre um relâmpago, é emitido energia em várias formas, incluindo luz visível e radiação infravermelha. O "evento de energia" é uma quantificação dessa energia. Caso não haja eventos na hora informada foi setado o valor para 0, logo não houve raio no momento.

Para a série temporal do TPW, observou-se uma quantidade significativa de valores faltantes nessa fonte de dados. Em vez de simplesmente excluir esses dados, optamos por utilizar uma coluna indicadora de ausência de dados. Essa coluna indicadora é uma variável binária que sinaliza se um valor correspondente em outra coluna está faltando. Essa abordagem transforma a falta de dados em um atributo a ser considerado. Para implementar isso nos dados de TPW, procedemos da seguinte forma: examinamos cada linha do conjunto de dados para determinar se o valor de TPW estava ausente. Se estivesse, atribuíamos 1 na nova coluna, se não, atribuíamos 0. Em relação à coluna TPW original, completamos as entradas com dados faltantes usando a média dos valores existentes.

datetime	TPW
2019-12-21 16:07:37	NaN
2019-12-22 11:07:42	NaN
2020-03-20 16:07:38	NaN
2020-03-21 11:02:41	NaN
...	...
2023-10-01 01:15:05	NaN
2023-10-01 01:25:05	44.347668
2023-10-01 01:35:05	44.006908
2023-10-01 01:45:05	45.185452
2023-10-01 01:55:05	45.266331

Tabela 4.8: TPW dataframe antes de merge e criação da nova coluna.

datetime	temperature	relative humidity	...	precipitation	TPW	tpw idx missing
2018-01-01 00:00:00	0.465798	0.795699	...	0.0	42.579010	1
2018-01-01 01:00:00	0.442997	0.849462	...	0.0	42.579010	1
2018-01-01 02:00:00	0.452769	0.860215	...	0.0	42.579010	1
2018-01-01 03:00:00	0.442997	0.870968	...	0.0	42.579010	1
2018-01-01 04:00:00	0.436482	0.870968	...	0.0	42.579010	1
...	...	...	...	...	...	...
2023-05-31 19:00:00	0.257329	0.822581	...	1.6	30.427313	1
2023-05-31 20:00:00	0.283388	0.731183	...	0.0	30.427313	1
2023-05-31 21:00:00	0.273616	0.897849	...	0.0	30.427313	1
2023-05-31 22:00:00	0.254072	0.930108	...	0.0	30.427313	1
2023-05-31 23:00:00	0.234528	0.790323	...	0.0	30.427313	1

Tabela 4.9: TPW dataframe depois do merge e criação da nova coluna.

4.6 Treinamento

Depois que os dados são preparados por todos os passos anteriores, e as características são extraídas, o modelo é treinado usando um algoritmo de aprendizado de máquina.

Avaliação do Modelo: Após o treinamento, o desempenho do modelo é avaliado usando métricas de avaliação, como acurácia, precisão, revocação, pontuação F1 ou área sob a curva característica de operação do receptor (ROC). Essas métricas fornecem insights sobre o quão bem o modelo generaliza para dados não vistos e quão equilibradas são suas previsões entre as duas classes.

Predição: Uma vez que o modelo é treinado e avaliado, ele pode ser usado para fazer previsões em dados novos e não vistos. O modelo recebe as características extraídas como entrada e gera um rótulo de classe previsto ou uma pontuação de probabilidade indicando a probabilidade de pertencer a uma determinada classe. O limiar para decidir o rótulo de classe pode ser definido com base nos requisitos específicos do problema.

Ao desenvolver modelos de previsão com base nos dados coletados, é crucial compreender e analisar a estratégia mais apropriada para o contexto da pesquisa. No caso da previsão de precipitação, há várias abordagens disponíveis para lidar com esse problema. Duas abordagens amplamente utilizadas em problemas de aprendizado de máquina são a classificação binária e a classificação ordinal. Ambas as abordagens serão consideradas e avaliadas para determinar o melhor modelo a ser adotado. Isso assegurará uma análise abrangente e a seleção da estratégia mais eficiente para atender aos objetivos da pesquisa.

4.7 Divisão dos dados em treino, validação, teste

Na construção de modelos de aprendizado de máquina, uma etapa crítica é a divisão do conjunto de dados em subconjuntos distintos para treinamento, validação e teste. Essa divisão é fundamental para avaliar a eficácia e a generalização do modelo. No contexto do nosso projeto, adotamos a seguinte divisão: 70% dos dados para treinamento, 20% para validação e 10% para teste. Vamos explorar o papel e a importância de cada um desses conjuntos:

Conjunto de Treinamento (70%): Este é o maior subconjunto e é utilizado para treinar o modelo. Com 70% dos dados disponíveis, o modelo tem a oportunidade de aprender variadas nuances e padrões presentes nos dados. Durante o treinamento, o modelo ajusta seus parâmetros e coeficientes para melhor se adaptar a esses dados. A maior proporção de dados de treinamento visa garantir que o modelo tenha exposição suficiente à diversidade e às características do conjunto de dados completo.

Conjunto de Validação (20%): Após o treinamento inicial, o modelo é testado no conjunto de validação. Este subconjunto não é utilizado durante o treinamento, o que permite uma avaliação imparcial do desempenho do modelo. Os 20% destinados à validação são significativos para ajustar hiperparâmetros, realizar seleção de características e diagnosticar problemas como overfitting (superajuste), onde o modelo se ajusta demais aos dados de treinamento e perde a capacidade de generalizar para novos dados.

Conjunto de Teste (10%): Por fim, o conjunto de teste é usado para avaliar a performance final do modelo. Este subconjunto é mantido separado e não é utilizado em nenhuma fase do treinamento ou validação. O objetivo é testar a capacidade do modelo de generalizar para dados completamente novos e não vistos anteriormente. A proporção de 10% para o conjunto de teste é suficiente para fornecer uma avaliação precisa e confiável da eficácia geral do modelo, sem comprometer a quantidade de dados disponíveis para treinamento e validação.

4.7.1 Classificação binária

No campo do aprendizado de máquina, uma abordagem comumente utilizada é a classificação binária. Um modelo classificador binário recebe dados de entrada e gera uma única saída, que corresponde à classe associada [Domingos, 2012]. Alguns exemplos típicos dessa abordagem incluem a classificação de pessoas como doentes ou não-doentes, a análise de aprovação de crédito e outros cenários similares.

No contexto da previsão de precipitação, um modelo classificador binário busca responder, com base em seu treinamento, se uma determinada configuração de características pode ser classificada como “chuva” ou “sem chuva”. No entanto, uma das limitações dessa abordagem é que ela atribui igual importância a todas as classes. Isso significa que, se o modelo cometer um erro ao classificar como “sem chuva”, não é possível distinguir se o erro foi pequeno, indicando uma chuva leve, ou se foi mais grave, indicando uma chuva intensa.

4.7.2 Classificação ordinal

Diferentemente da abordagem anterior, em determinados problemas práticos de aprendizado de máquina, as classes apresentam uma natureza ordinal. Um exemplo comum de classificação ordinal pode ser observado no domínio da avaliação de desempenho de funcionários em uma organização. Ao realizar a avaliação, os gestores classificam os funcionários em categorias como “desempenho abaixo das expectativas”, “desempenho satisfatório”, “desempenho acima da média” ou “desempenho excepcional”.

Em certos contextos, um modelo que leve em consideração a ordem das classes pode ser preferível à abordagem anterior, uma vez que os algoritmos de classificação convencionais não conseguem aproveitar essa informação de ordem, tratando o atributo de classe como uma quantidade nominal. No caso da previsão de precipitação, essa abordagem poderia classificar os níveis de chuva em uma escala ordenada, variando desde “chuva leve” até “chuva intensa”. Para o desenvolvimento deste estudo específico, foram consideradas as classes definidas pelo Alerta-Rio. A Tabela 4.10 a seguir apresenta os diferentes níveis de chuva e suas respectivas medições de precipitação associadas em milímetros por hora.

Classificação	**Nível de precipitação (ρ) em mm/h**
Sem chuva	$\rho = 0$
Chuva fraca	$\rho \leq 5$
Chuva moderada	$5 < \rho \leq 25$
Chuva forte	$25 < \rho \leq 50$
Chuva extrema	$50 < \rho$

Tabela 4.10: Níveis de precipitação definidos pelo alerta-rio.

4.8 Avaliação dos modelos

Para avaliar os modelos gerados pelas abordagens de classificação, utilizamos tanto a métrica F1-Score quanto o Coeficiente de Matthews. Essas métricas são aplicadas para cada um dos níveis de precipitação e para o modelo como um todo. O F1-Score é calculado com base em precisão (P) e revocação (R) e foi utilizado para a Classificação Ordinal. O Coeficiente de Matthews é uma medida mais robusta que leva em conta todas as quatro categorias da matriz de confusão: verdadeiros positivos (TP), falsos positivos (FP), verdadeiros negativos (TN) e falsos negativos (FN) e foi utilizado para a Classificação Binária.

O F1-Score é expresso pelas Equações 4.4, 4.5 e 4.6, onde T_p representa os verdadeiros positivos (valores previstos corretamente para a ocorrência de chuva), F_p os falsos positivos (valores previstos incorretamente para a ocorrência de chuva) e F_n os falsos negativos (valores previstos incorretamente para a não ocorrência de chuva).

$$P = \frac{T_p}{T_p + F_p} \tag{4.4}$$

$$R = \frac{T_p}{T_p + F_n} \tag{4.5}$$

$$F_1 = 2 * \frac{P * R}{P + R} \tag{4.6}$$

O Coeficiente de Matthews é calculado usando a Equação 4.7:

$$MCC = \frac{T_p \times T_n - F_p \times F_n}{\sqrt{(T_p + F_p)(T_p + F_n)(T_n + F_p)(T_n + F_n)}} \tag{4.7}$$

Neste caso, T_n representa os verdadeiros negativos (valores previstos corretamente para a não ocorrência de chuva). O Coeficiente de Matthews é particularmente útil em situações com desequilíbrio de classes, oferecendo uma perspectiva mais equilibrada sobre o desempenho do modelo.

Utilizamos a métrica F1 por meio da função *classification_report* da biblioteca Scikit-Learn. Esta função gera um relatório que avalia a qualidade preditiva do modelo ao fornecer dois conjuntos de valores, um contendo os dados esperados e o outro os valores de precipitação previstos pelo modelo. Para o Coeficiente de Matthews, também utilizamos recursos do Scikit-Learn,

especificamente a função *matthews_corrcoef*, que calcula o coeficiente a partir dos valores verdadeiros e dos valores previstos.

Capítulo 5

Avaliação experimental

No presente capítulo, trazemos uma descrição dos experimentos conduzidos e dos resultados obtidos. Na Seção 5.1, descrevemos como os conjuntos de dados foram gerados para treinamento, validação e teste. Na seção 5.2 são descritos os hiperparâmetros utilizados na construção da ANN. Na Seção 5.3, apresentamos os resultados obtidos, e realizamos uma discussão sobre eles.

5.1 Conjunto de dados

Nessa seção, descrevemos como os conjuntos de dados foram gerados para treinamento, validação e teste, tanto para apenas os dados do INMET quanto os do INMET com adição dos dados do satélite.

Essa divisão em conjuntos distintos tem o objetivo de garantir uma avaliação justa e imparcial do desempenho do modelo de classificação. O conjunto de treinamento é utilizado para aprender os padrões e características dos dados, o conjunto de validação é empregado para ajustar os hiperparâmetros e o conjunto de teste é reservado para avaliar a capacidade de generalização do modelo em dados não vistos durante o treinamento.

É importante ressaltar que a escolha das proporções para a divisão dos conjuntos de treinamento, validação e teste pode variar dependendo do problema e das características dos dados. A proporção mencionada neste trabalho foi adotada como uma abordagem comum, mas outras divisões podem ser utilizadas de acordo com a natureza específica do problema em questão.

Classificação binária. Para somente os dados do INMET, utilizamos um total de 36664 dados. A divisão dos conjuntos foi feita da seguinte maneira: 70% para treinamento, 20% para validação e 10% para teste. O conjunto de treinamento, foi utilizado para ajustar os parâmetros do modelo durante o processo de treinamento. O conjunto de validação, foi usado para ajustar os hiperparâmetros do modelo e evitar o overfitting. Por fim, o conjunto de teste, foi reservado para avaliar o desempenho final do modelo.

Classificação ordinal. Assim como anteriormente, para a classificação ordinal usamos o mesmo conjunto de dados com a mesma quantidade de exemplos: 36664. Dividido nos mesmos percentuais, ficamos com 70% dos dados para treinamento, 20% dos dados para validação e 10% dos dados para teste.

5.2 Construção das redes neurais

Nesta seção, descrevemos a rede e os hiperparâmetros utilizados na construção da ANN para o modelo de Classificação Binária e o modelo de Classificação Ordinal. Esses hiperparâmetros foram predeterminados no escopo desse trabalho. Em todos os experimentos, usamos a linguagem Python versão 3.8. Realizamos esses experimentos em uma máquina com sistema operacional Windows 11, processador *AMD Ryzen 5 3600* e memória *RAM* de 24 GB.

5.2.1 Estrutura da Rede Neural

A rede implementada por nós no escopo deste trabalho é composta por uma sequência de operações que processam dados sequenciais unidimensionais. Ela utiliza uma combinação de camadas convolucionais e totalmente conectadas para executar tarefas de classificação. A rede pode ser dividida em duas partes principais: o extrator de características e o classificador.

Extrator de Características

O extrator de características é responsável por transformar o sinal de entrada bruto em uma representação de alto nível que destila a informação essencial necessária para a classificação.

A Camada Convolucional 1D: Esta camada aplica uma série de filtros convolucionais aos dados de entrada. Cada filtro desliza ao longo da sequência de entrada (a dimensão temporal) e produz uma saída que é uma combinação linear dos valores dentro de sua janela de visão. A operação convolucional captura padrões locais dentro da sequência, tais como tendências, picos, ou quedas no contexto de séries temporais. O padding adicionado garante que a saída da convolução tenha o mesmo comprimento que a entrada, mantendo assim a resolução temporal.

Função de Ativação ReLU: Após a convolução, uma função de ativação ReLU é aplicada para introduzir não-linearidades no modelo. A ReLU tem a vantagem de ser computacionalmente eficiente e reduzir o problema de gradientes que desaparecem durante o treinamento.

A função de ativação ReLU, ou Unidade Linear Retificada, é uma função comumente usada

em redes neurais que é definida matematicamente como:

$$ReLU(x) = max(0,x). \tag{5.1}$$

O que isso significa é simples: para qualquer valor de entrada x, a função ReLU produz x se x for positivo e produz 0 se x for negativo ou zero. Visualmente, a função se parece com um gráfico que aumenta linearmente para valores de entrada positivos e permanece em 0 para valores negativos.

Regularização Dropout: Para evitar overfitting, um dropout é aplicado após a ativação ReLU. Esta técnica envolve a "desativação" aleatória de alguns neurônios durante o treinamento, forçando a rede a aprender representações mais robustas dos dados.

Classificador

O classificador pega a representação de características de alto nível do extrator e a transforma em uma previsão de classe final.

Camada Linear: A primeira camada totalmente conectada, ou densa, recebe as características achatadas do extrator como entrada e as transforma em um espaço de 50 dimensões. Esta camada permite que a rede aprenda combinações lineares das características para formar representações ainda mais abstratas.

Segunda Função de Ativação ReLU: Uma segunda ativação ReLU segue a camada totalmente conectada para adicionar outra camada de não-linearidade e capacidade de aprendizado.

Segunda Camada Linear: Uma segunda camada linear reduz a dimensão de 50 para o número final de saídas desejadas. O número de saídas corresponde ao número de classes que o modelo está tentando prever.

Função de Ativação Sigmoid: Finalmente, uma função de ativação sigmoid é aplicada para converter as saídas lineares em probabilidades entre 0 e 1. Em um problema de classificação binária, isso pode ser interpretado como a probabilidade de a entrada pertencer a uma das duas classes.

A função de ativação sigmoid é uma função matemática que mapeia qualquer valor de entrada real para um valor de saída no intervalo (0, 1). É definida pela fórmula:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{5.2}$$

- Onde $\sigma(x)$ é o valor de saída da função sigmoide para a entrada x,
- *e* é a base do logaritmo natural.

Dimensionamento e Formato dos Dados

Um aspecto importante da arquitetura é a manipulação das dimensões dos dados de entrada para se adequar às expectativas das camadas da rede. Em particular, a rede espera que os dados de entrada para a camada convolucional estejam na forma, o que pode exigir a reorganização das dimensões dos dados antes de alimentá-los na rede.

5.2.2 Hiperparâmetros

Taxa de aprendizado (Learning Rate)

A taxa de aprendizado controla o tamanho dos passos que o otimizador dará ao atualizar os pesos da rede durante o treinamento. Para o modelo de Classificação Binária, a taxa de aprendizado foi definida como 0,0003. Já para o modelo de Classificação Ordinal, utilizamos uma taxa de aprendizado de 3,0e-06.

Quantidade de épocas de treinamento (Number of Epochs)

Uma época de treinamento ocorre quando todos os exemplos do conjunto de treinamento passam pela rede neural uma vez. Para o modelo de Classificação Binária, foram utilizadas 3500 épocas de treinamento. Para o modelo de Classificação Ordinal, foram utilizadas 8000 épocas de treinamento.

Tamanho do lote (Batch Size)

O tamanho do lote refere-se à quantidade de exemplos de treinamento apresentados à rede antes de calcular os gradientes e atualizar os pesos. Nesse caso, tanto para o modelo de Classificação Binária quanto para o modelo de Classificação Ordinal, um tamanho de lote de 1024 foi escolhido.

Regularização de pesos (Weight Decay)

A regularização de pesos é uma técnica usada para evitar o overfitting (sobreajuste) durante o treinamento da rede neural. Para ambos as redes neurais, um valor de regularização de peso de 0 foi utilizado, indicando a ausência dessa regularização.

Taxa de dropout (Dropout Rate)

O dropout é uma técnica de regularização que visa evitar o overfitting ao desativar aleatoriamente unidades (neurônios) durante o treinamento. Para ambas as redes neurais foi aplicada uma taxa de dropout de 0,5, indicando que cada neurônio tinha uma probabilidade de 50% de ser desativado a cada passo de treinamento.

5.3 Resultados

Nesta seção, fornecemos uma análise dos resultados dos experimentos conduzidos para os modelos de previsão. Para avaliar o desempenho dos modelos para cada estação foram adotadas as métricas F1-Score e Coeficiente de Matthews para Classificação Ordinal e Classificação Binária, respectivamente. Quanto mais próximo de 1 melhor a previsão. Além disso, para combinar os dados provenientes do INMET com os dados do satélite GOES-16, foram definidas abreviações específicas, conforme descrito a seguir:

Código da Estação	Nome da Estação
A602	Marambaia
A621	Vila militar
A627	Niterói
A636	Jacarepaguá
A652	Forte de Copacabana

Tabela 5.1: Estações meteorológicas e seus nomes correspondentes.

Classificação Binária

Para a Classificação Binária, percebemos que a inclusão de fontes de dados do satélite (seja do produto GLM ou do produto TPW) melhora, em alguns casos, a previsão.

A Figura 5.1 apresenta a comparação do Coeficiente de Matthews por estação, conforme podemos ver a inclusão de outras fontes de dados mostra um melhora em alguns casos.

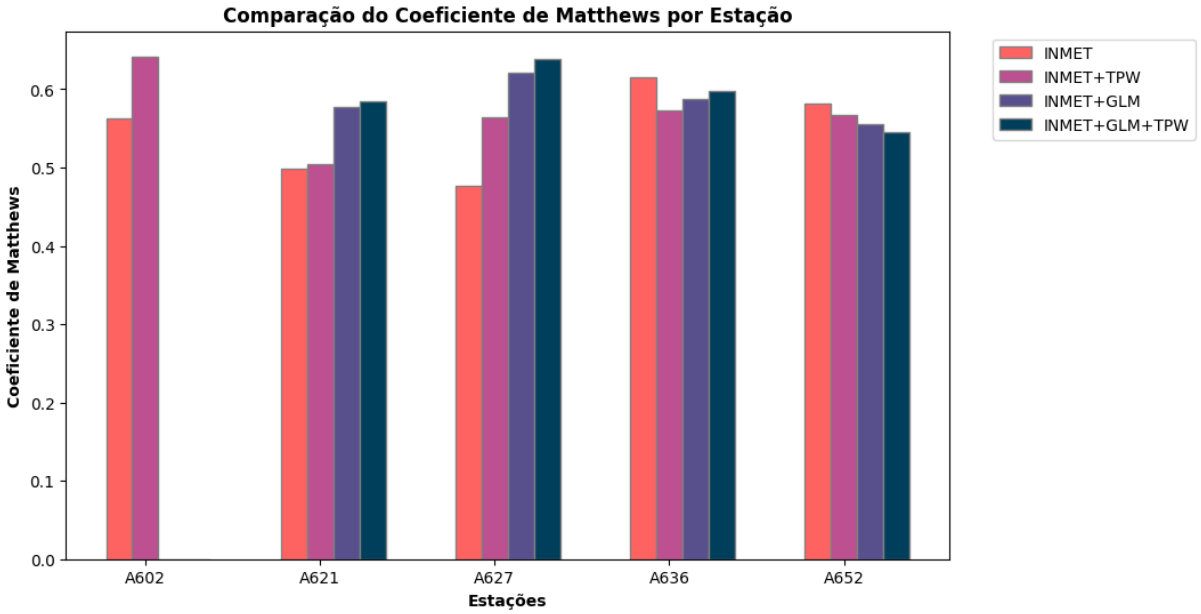

Figura 5.1: Coeficiente de Matthews para cada combinação por estação.

Para a Estação A602, a inclusão dos dados TPW melhorou bastante a previsão. Para as Estações A621 e A627 a combinação dos dados do INMET com os dados do GLM e do TPW gerou o melhor resultado. Já para as Estações A636 e A652, as diferentes combinações de fontes de dados não gerou uma melhoria na previsão, na verdade as piorou, evidenciando que nem sempre a inclusão de tais dados é benéfica para a previsão.

Classificação Ordinal

Para a Classificação Ordinal, iremos mostrar as diferentes combinações de dados para cada Estação de Interesse. Vale notar que, conforme evidenciado na Seção 4.4, algumas estações tinham poucos, ou nenhum, dados referentes a alguns níveis de chuva.

A Figura 5.2 mostra a comparação do F1-Score por Classe para a Estação A602. Para a Estação A602, onde tínhamos apenas dados do produto TPW, a inclusão do mesmo não melhorou a previsão de dias sem chuva, porém ajudou o modelo a detectar e prever dias de chuva fraca.

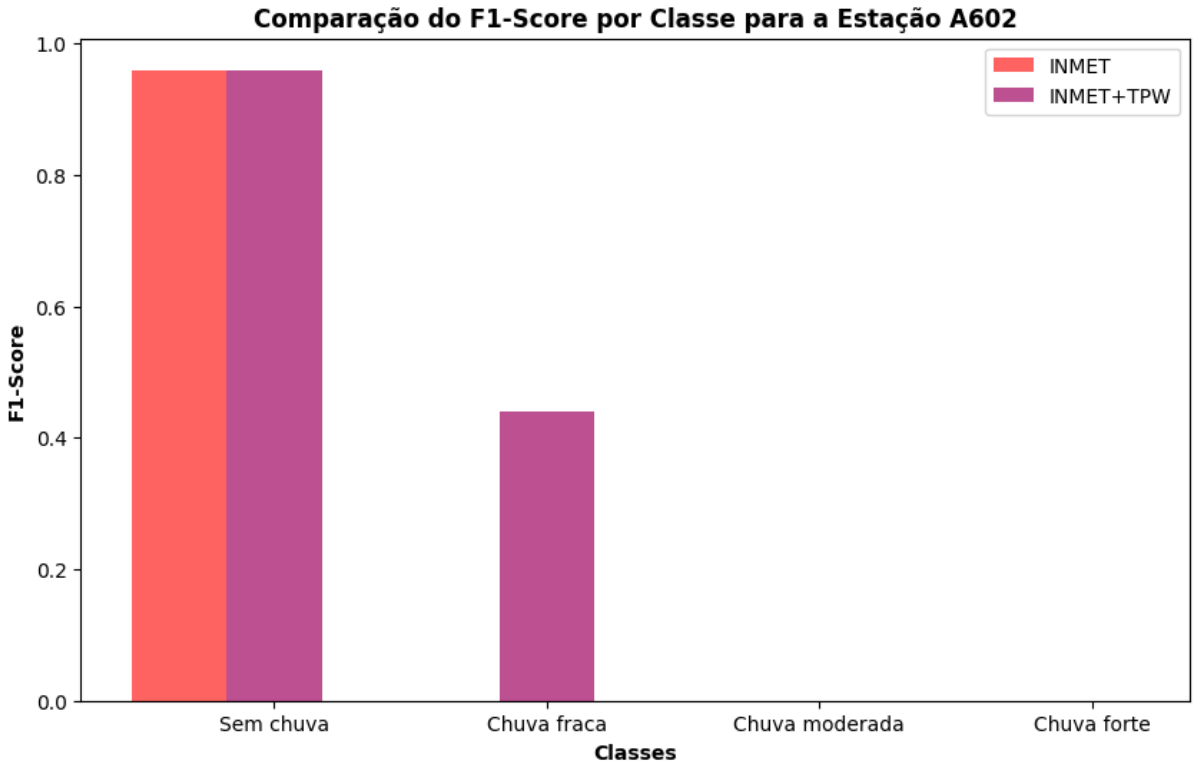

Figura 5.2: F1-Score das diferentes combinações de dados para cada nível de chuva na Estação A602.

A Figura 5.3 mostra a comparação do F1-Score por Classe para a Estação A621. Para a Estação A621, a previsão de não chuva piorou com cada diferente combinação de dados. Já para a previsão de chuva fraca, a combinação de todos os dados foi o que gerou o melhor resultado, conforme visto na Figura.

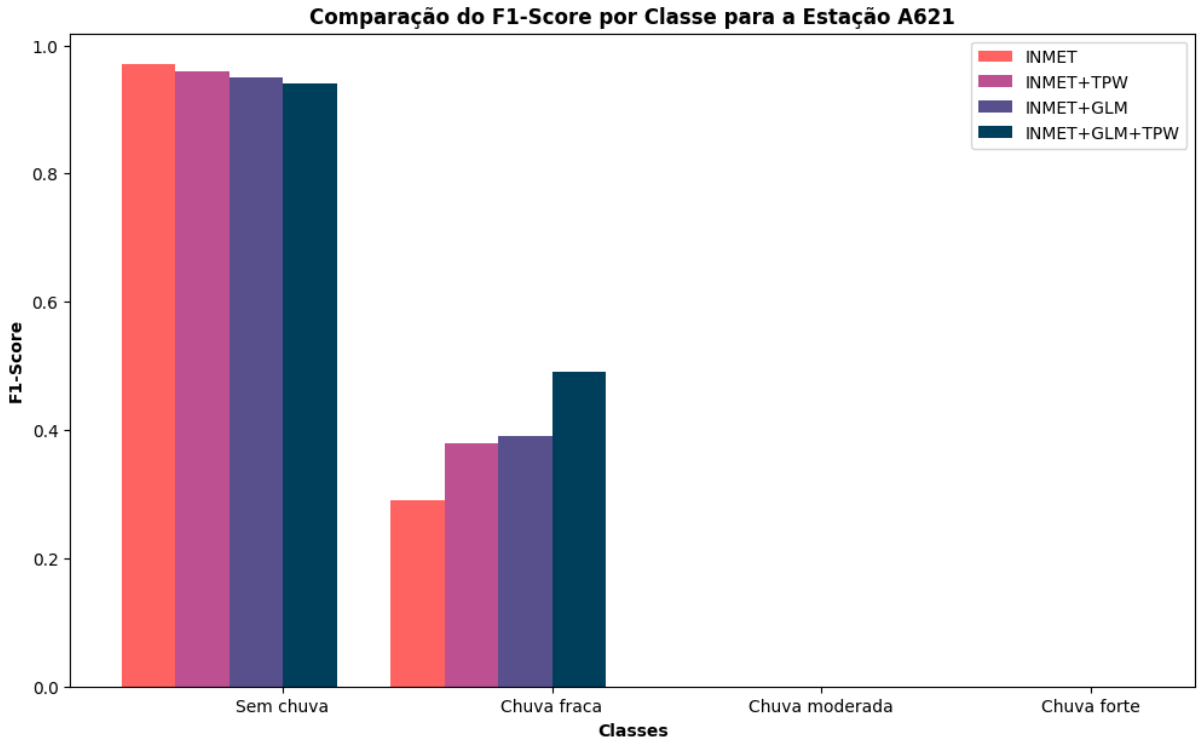

Figura 5.3: F1-Score das diferentes combinações de dados para cada nível de chuva na Estação A621.

A Figura 5.4 apresenta a comparação do F1-Score por Classe para a Estação A627. Para a Estação A627, a previsão de não chuva se manteve bastante estável com as diferentes combinações de dados, sendo a melhor a combinação dos dados do INMET e do TPW. Para dias de chuva fraca, a inclusão individual de dados do GLM ou TPW piorou a previsão, mas quando combinamos ambos com os dados do INMET tivemos um ganho considerável na previsão.

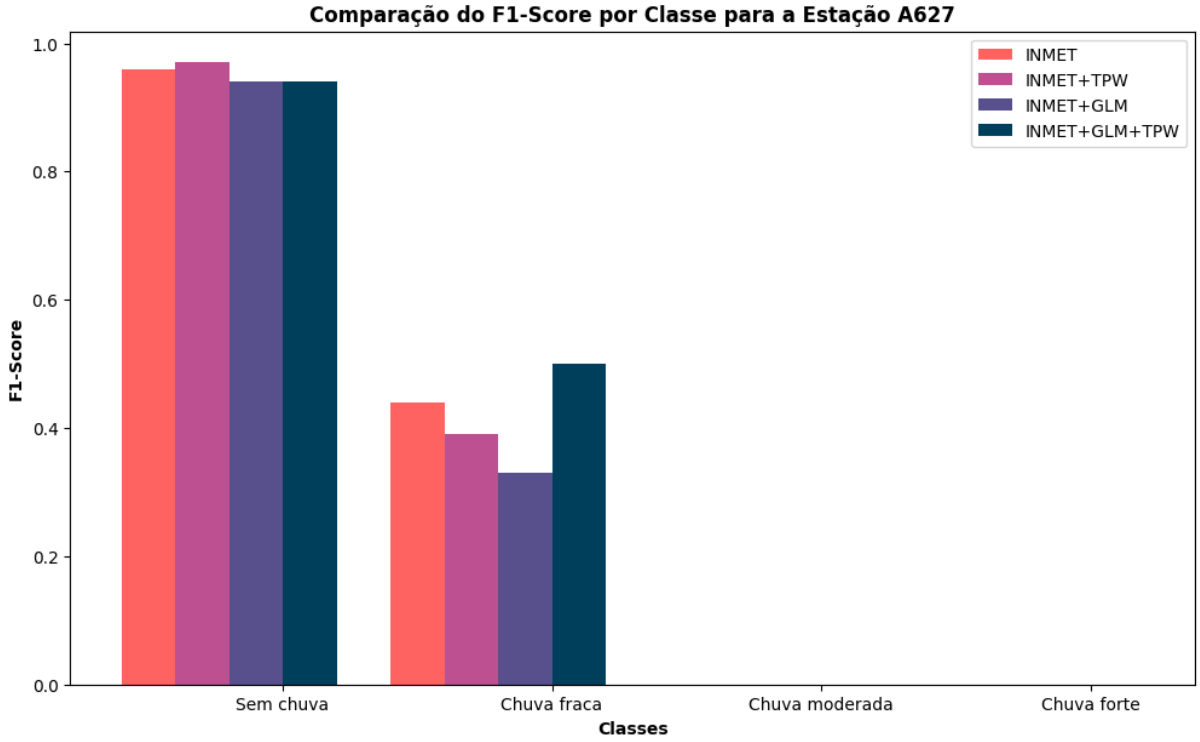

Figura 5.4: F1-Score das diferentes combinações de dados para cada nível de chuva na Estação A627.

A Figura 5.5 apresenta a comparação do F1-Score por Classe para a Estação A636. Para a Estação A636, a previsão de dias sem chuva não melhorou com nenhuma combinação de dados, porém a utilização de dados do INMET junto dos dados provenientes do TPW gerou o mesmo resultado que utilizando os dados do INMET sozinhos. Para a previsão de chuva fraca, os dados do TPW se mostraram ineficazes, porém ao combinar os dados do INMET com os dados do GLM a nossa previsão melhora bastante.

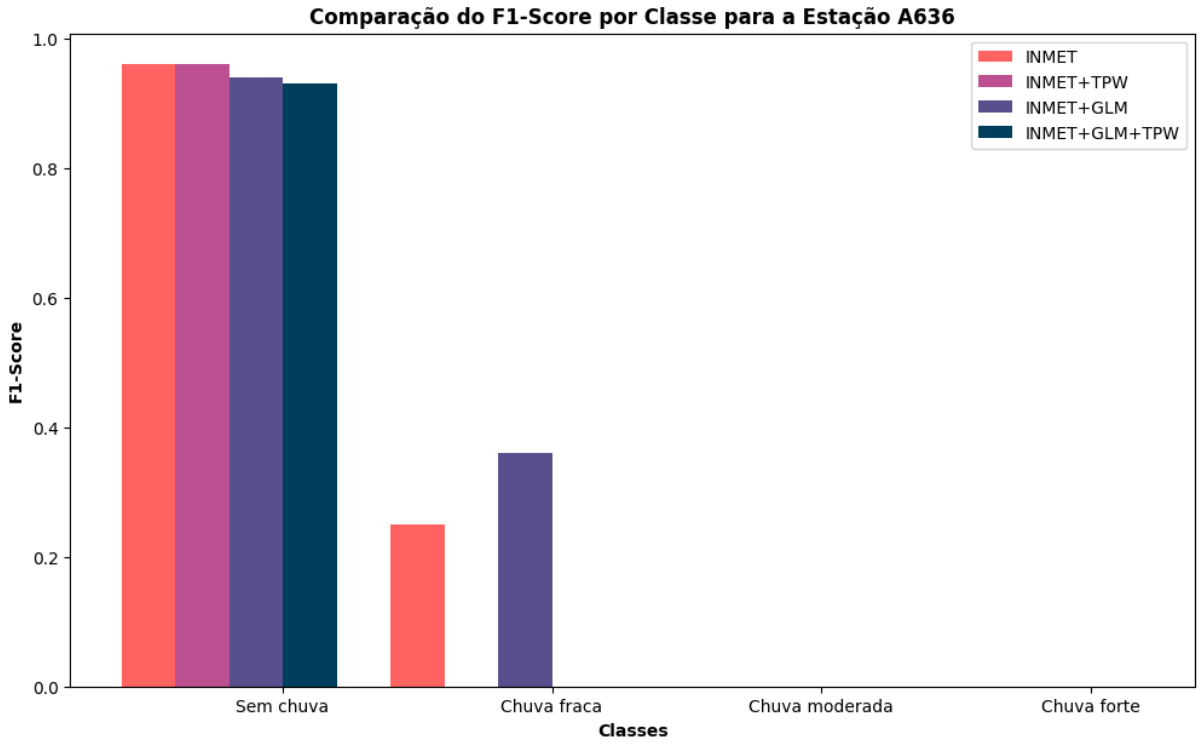

Figura 5.5: F1-Score das diferentes combinações de dados para cada nível de chuva na Estação A636.

A Figura 5.6 apresenta o F1-Score para a Estação A652. A previsão de dias sem chuva melhorou um pouco quando utilizado apenas os dados do INMET com os dados do TPW. Para a previsão de chuva fraca, nenhuma combinação foi capaz de gerar resultados, apenas os dados originais do INMET.

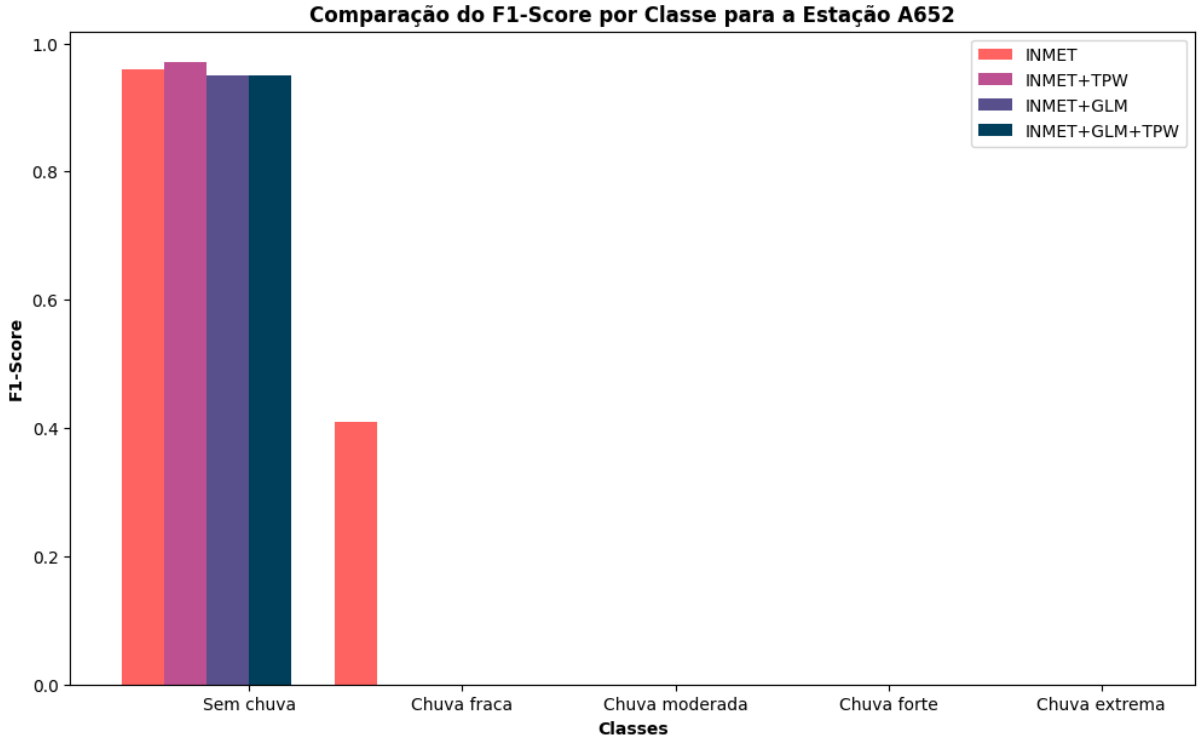

Figura 5.6: F1-Score das diferentes combinações de dados para cada nível de chuva na Estação A652.

5.3.1 Discussão

A análise dos dados adquiridos das diversas estações meteorológicas e a incorporação de dados de satélite revela *insights* cruciais sobre o desempenho dos modelos de classificação tanto binária quanto ordinal. Um dos aspectos mais notáveis é a disparidade entre a alta acurácia e precisão para a classe majoritária (classe 0), em comparação com a classe minoritária (classe 1), o que é consistente em quase todos os modelos. Isso pode indicar um viés do modelo favorecendo a classe majoritária, o que é uma ocorrência comum em conjuntos de dados desequilibrados. O fenômeno é ainda mais acentuado nos modelos ordinais, onde as classes 2 e 3, provavelmente ainda menos representadas, têm um desempenho ainda pior.

Quando consideramos a fonte de dados, os modelos que utilizam exclusivamente os dados da estação já apresentam um desempenho bastante alto. Isso sugere que os dados do INMET são robustos e informativos o suficiente para permitir previsões precisas. Contudo, a introdução de dados do produto GLM não melhorou o desempenho dos modelos significativamente. Uma possibilidade é que os dados do GLM, apesar de adicionarem uma dimensão adicional de informação, podem introduzir ruído ou não serem tão relevantes para a tarefa específica de previsão

meteorológica.

Por outro lado, a inclusão de dados do produto TPW melhorou o desempenho em determinadas estações, o que sugere que essas informações adicionais fornecem insights valiosos que auxiliam na previsão. A melhoria do desempenho com a adição de dados TPW em relação à adição de dados GLM pode indicar que o conteúdo de água precipitável é um preditor mais forte ou mais relevante para os fenômenos que os modelos estão tentando prever.

Interessantemente, a combinação de ambos os conjuntos de dados do satélite não produziu os ganhos esperados na precisão ou acurácia. Isso pode ser devido à complexidade adicional nos dados não ser adequadamente capturada pelos modelos atuais, ou pode ser que os modelos estejam encontrando dificuldades em integrar essas fontes de dados de maneira a produzir previsões melhoradas. Uma investigação mais aprofundada sobre como essas fontes de dados interagem entre si e com o modelo poderia revelar por que a adição de mais informações nem sempre se traduz em melhor desempenho.

Outro ponto de discussão é a consistência da estação Forte de Copacabana em apresentar bons resultados, o que pode ser atribuído a fatores como localização geográfica, condições climáticas locais mais homogêneas ou melhores instrumentos de medição. Essa consistência aponta para a importância da qualidade dos dados na modelagem preditiva.

A discussão desses resultados pode beneficiar-se de uma exploração mais aprofundada dos modelos, como análise de características, importância das variáveis e ajuste de hiperparâmetros. Além disso, poderia ser útil investigar métodos para lidar com o desequilíbrio das classes, como a geração de dados sintéticos para as classes minoritárias ou técnicas de ponderação de classes. Estas abordagens poderiam potencialmente melhorar a precisão e o recall para as classes minoritárias, levando a um modelo mais equilibrado e justamente avaliado.

Capítulo 6

Considerações finais

Neste capítulo de conclusão, realizaremos uma análise abrangente dos resultados desta pesquisa, destacando sua relevância para o tema em estudo. Apresentaremos um resumo dos principais achados, demonstrando como eles estão alinhados com os objetivos estabelecidos. Além disso, abordaremos as limitações identificadas durante o estudo.

6.1 Análise retrospectiva

Os experimentos realizados demonstraram que modelos de classificação aplicados a dados de estações meteorológicas individuais oferecem previsões acuradas, mas enfrentam desafios ao classificar eventos menos frequentes, sugerindo a necessidade de técnicas avançadas para lidar com desequilíbrios de classes. A inclusão de dados de satélite TPW melhorou o desempenho dos modelos, enquanto a adição de múltiplos conjuntos de dados não aumentou a eficácia proporcionalmente, indicando que a integração e a seleção criteriosa de características são cruciais. Este estudo sublinha a importância da qualidade dos dados e oferece direções para futuras pesquisas na melhoria de modelos preditivos em meteorologia, com implicações práticas para previsões mais confiáveis.

Uma dificuldade encontrada nos experimentos foi o fato das séries históricas que utilizamos como base não possuírem muitos valores de chuva forte e chuva extrema no período de teste. Isso dificulta a capacidade dos modelos de realizarem previsões dessas categorias específicas, por não ter amostras o bastante no seu período de treinamento e validação. Essa tendência no conjunto de dados impacta negativamente na precisão do modelo, para previsões de eventos de alta magnitude. Outra dificuldade encontrada foi no tratamento dos dados fornecidos pelo satélite, onde demandava muito tempo para o download e pre processamento dos mesmos.

6.2 Trabalhos futuros

Os resultados aqui expostos nos mostram que ao utilizar CNN 1D com a presença de uma nova fonte de dados não necessariamente gera melhoras nas previsões de chuva, porém ainda há alguns pontos que podem ser vistos mais adiante.

Como citado ao longo do trabalho, o satélite GOES-16 possui diversos produtos, dentre os quais foi escolhido o GLM e TPW para ser o nosso alvo de estudo nesse momento. Com essa vasta gama de opções é natural se pensar em acrescentar novos produtos na continuação do trabalho. Um dos produtos que se destaca é o ABI, onde dentre eles estamos utilizando o TPW, porem temos uma enorme variedade de tipos além do TPW, que fornecem imagens visuais e térmicas da Terra em intervalos de 5 minutos, e pode ser um dos alvos de experimentos no futuro desse trabalho.

Um estudo sobre a busca dos melhores hiperparâmetros também se mostra necessário dado que, para o escopo desse trabalho, utilizamos valores predeterminados e não investigamos para tentar obter os melhores valores possíveis.

Conforme visto nas tabelas na seção de resultados, os dados estão muito desbalanceados, o que abre espaço para estudo de técnicas de oversampling. Essa técnica que consiste, basicamente, em aumentar a quantidade de registros da classe com menor frequência até que a base de dados possua uma quantidade equilibrada entre as classes da variável alvo.

Outra adição que pode ser estudada é a inclusão de outro modelo diferente do CNN 1D. A ideia é buscarmos, juntos a esse possível novo modelo, alcançar melhorias consideráveis nas previsões em curto prazo de chuva.

Referências Bibliográficas

Shreya Agrawal, Luke Barrington, Carla Bromberg, John Burge, Cenk Gazen, and Jason Hickey. Machine learning for precipitation nowcasting from radar images. *arXiv preprint arXiv:1912.12132*, 2019. 25

Marta Banbura, Domenico Giannone, and Lucrezia Reichlin. Nowcasting. 2010. 6

Paolo Bonelli and Pietro Marcacci. Thunderstorm nowcasting by means of lightning and radar data: algorithms and applications in northern italy. *Natural Hazards and Earth System Sciences*, 8(5):1187–1198, 2008. 24

Cooperative Institute for Meteorological Satellite Studies CIMSS. URL `http://cimss.ssec.wisc.edu/satmet/modules/1_intro/intro-1.html`. 7, 8, 9

Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, oct 2012. ISSN 0001-0782. doi: 10.1145/2347736.2347755. 37

e2eml, 2020. URL `https://e2eml.school/convolution_one_d.html`. 21

EUMeTrain. URL `https://resources.eumetrain.org/data/3/359/navmenu.php?tab=2&page=1.0.0`. 12

Carme Farnell, Tomeu Rigo, and Nicolau Pineda. Lightning jump as a nowcast predictor: Application to severe weather events in catalonia. *Atmospheric Research*, 183:130–141, 2017. 25

Artur Gonçalves Ferreira. *Meteorologia prática*. Oficina de Textos, 2006. 8

Teresa Gallotti Florenzano. *Os satélites e suas aplicações*. Sindicato dos servidores Públicos Federais na Área de Ciências e Tecnologia . . . , 2008. 8

Lei Han, Juanzhen Sun, Wei Zhang, Yuanyuan Xiu, Hailei Feng, and Yinjing Lin. A machine learning nowcasting method based on real-time reanalysis data. *Journal of Geophysical Research*, 122(7):4038–4051, 2017. doi: 10.1002/2016JD025783. 2

Ministério da Agricultura INMET. URL `https://portal.inmet.gov.br/sobre-meteorologia`. 6, 7

Athanasios Karagiannidis, Konstantinos Lagouvardos, and Vassiliki Kotroni. The use of lightning data and meteosat infrared imagery for the nowcasting of lightning activity. *Atmospheric Research*, 168:57–69, 2016. 25

Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J Inman. 1d convolutional neural networks and applications: A survey. *Mechanical systems and signal processing*, 151:107398, 2021. x, 19, 20

Moriah Kohn, Eli Galanti, Colin Price, Kostas Lagouvardos, and Vassiliki Kotroni. Nowcasting thunderstorms in the mediterranean region using lightning data. *Atmospheric Research*, 100 (4):489–502, 2011. 24

Ryan Lagerquist, Amy McGovern, and Travis Smith. Machine learning for real-time prediction of damaging straight-line convective wind. *Weather and Forecasting*, 32(6):2175–2193, 2017. doi: 10.1175/WAF-D-17-0038.1. 2

Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436 – 444, 2015. doi: 10.1038/nature14539. Type: Review. 2

Hagyeong Lee and Jongwoo Song. Introduction to convolutional neural network using Keras; An understanding from a statistician. *Communications for Statistical Applications and Methods*, 26(6):591 – 610, 2019. doi: 10.29220/CSAM.2019.26.6.591. Type: Article. 2

Jussi Leinonen, Ulrich Hamann, Urs Germann, and John R Mecikalski. Nowcasting thunderstorm hazards using machine learning: the impact of data sources on performance. *Natural Hazards and Earth System Sciences*, 22(2):577–597, 2022. 26

José Antônio Marengo. Água e mudanças climáticas. *Estudos avançados*, 22:83–96, 2008. 1

NASA. URL `https://ghrc.nsstc.nasa.gov/lightning/overview_glm.html`. x, 14, 15

Alfredo Nespoli, Alessandro Niccolai, Emanuele Ogliari, Giovanni Perego, Elena Collino, and Dario Ronzio. Machine learning techniques for solar irradiation nowcasting: Cloud type classification forecast through satellite data and imagery. *Applied Energy*, 305:117834, 2022. 26

NOAA, a. URL `https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C01515`. 12, 14

NOAA and NASA. URL https://www.goes-r.gov/multimedia/videosForKids.html. 10

NASA NOAA, b. URL https://www.goes-r.gov/. 8

R. Rew and G. Davis. Netcdf: an interface for scientific data access. *IEEE Computer Graphics and Applications*, 10(4):76–82, 1990. doi: 10.1109/38.56302. 16

Franziska Schmid, Yong Wang, and Abdoulaye Harou. Nowcasting guidelines–a summary. *Bulletin*, 68:2, 2019. 7

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 25

Israel da Silva Filho et al. O uso de imagens de satélite no ensino de geografia: possibilidades e limitações na educação básica. 2008. 7

Ravisutha Sakrepatna Srinivasamurthy. *Understanding 1D Convolutional Neural Networks Using Multiclass Time-Varying Signalss*. PhD thesis, Clemson University, 2018. 19

Nicholas Stern. Stern review: the economics of climate change, Oct 2006. 1

Leonard Uhr and Charles Vossler. A pattern recognition program that generates, evaluates, and adjusts its own operators. In *Proceedings of the Western Joint Computer Conference: Extending Man's Intellect, IRE-AIEE-ACM 1961*, pages 555–569, 1961. doi: 10.1145/1460690.1460751. 2

James W. Wilson, Yerong Feng, Min Chen, and Rita D. Roberts. Nowcasting challenges during the Beijing olympics: Successes, failures, and implications for future nowcasting systems. *Weather and Forecasting*, 25(6):1691–1714, 2010. doi: 10.1175/2010WAF2222417.1. 2

Kanghui Zhou, Yongguang Zheng, Wansheng Dong, and Tingbo Wang. A deep learning network for cloud-to-ground lightning nowcasting with multisource data. *Journal of Atmospheric and Oceanic Technology*, 37(5):927–942, 2020. 25